Universidad
Carlos III de Madrid

TESIS DOCTORAL

Alta Precisión Relativa en Problemas de Álgebra Lineal Numérica en Matrices con Estructura

Autor:
Johan Armando Ceballos Cañón

Director:
Juan Manuel Molera Molera

DEPARTAMENTO DE MATEMÁTICAS

Leganés, Julio de 2013

TESIS DOCTORAL

Alta Precisión Relativa en Problemas de Álgebra Lineal Numérica en Matrices con Estructura

Autor: Johan Armando Ceballos Cañón

Director: Juan Manuel Molera Molera

Firma del Tribunal Calificador:

Firma

Presidente: José Javier Martínez Fernández de las Heras

Vocal: Esther Dopazo González

Secretario: Fernando de Terán Vergara

Calificación:

Leganés, de de

A mi hija Mariana

Este trabajo se desarrolló en el Departamento de Matemáticas de la Universidad Carlos III de Madrid Campus de Leganés bajo la dirección del profesor Juan Manuel Molera Molera. Se contó con una beca de la UC3M para el periodo del Máster y posteriormente de un contrato de Personal Investigador de la Comunidad de Madrid. Adicionalmente se recibió ayuda parcial de los siguientes proyectos de investigación: MINISTERIO DE EDUCACIÓN Y CIENCIA DIRECCIÓN GENERAL DE INVESTIGACIÓN 2006/03964/001, MINISTERIO DE CIENCIA E INNOVACIÓN 2010/00014/001, COMUNIDAD DE MADRID - UC3M 2011/00120/001 y MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD 2013/00065/001.

Agradecimientos

El principal y más merecido agradecimiento se lo debo a mi director de tesis, Juan Manuel Molera, por haberme introducido y guiado en el mundo de la investigación lo cual hizo posible esta tesis. También quiero expresar mi agradecimiento a mis compañeros del grupo de Algebra Lineal Numérica, en especial a Froilán M. Dopico por haberme permitido formar parte del mismo, de los proyectos de investigación, por sus consejos y charlas que fueron vitales para el trabajo; sin duda aprendí mucho gracias a ellos. Debo extender este agradecimiento al Departamento de Matemáticas por el apoyo financiero, a su planta de profesores y en especial al personal administrativo: Natalia Delgado y Alberto Calvo por su apoyo y buena disposición. A mi gran amigo Javier Pérez que siempre estuvo ahí para brindarme una palabra de aliento en el momento preciso.

Por último, quisiera expresar mi más profundo agradecimiento a mi hija Mariana y a mis padres por su constante cariño, confianza y apoyo, a ellos quiero dedicar esta tesis.

Índice general

Agradecimientos iii

Notación ix

Resumen xi

1. Introducción 1

2. RRD y HRA: Preliminares y resultados previos 17

2.1. Notación 17

2.2. Teoría de perturbaciones de EVP y SVD 18

2.3. Teoría de perturbaciones de SEL y LSP 20

2.4. Descomposiciones que revelan el rango y teoría de perturbaciones . . 23

2.4.1. Descomposición en valores singulares y problema simétrico de autovalores 24

2.4.2. Sistemas de ecuaciones lineales vía la RRD 26

2.5. Algoritmos y errores 27

3. Teoría de perturbaciones multiplicativas y soluciones precisas del LSP 35

3.1. Preliminares y conceptos básicos 38

3.2. Perturbaciones multiplicativas para la pseudoinversa de Moore-Penrose 40
3.2.1. Cotas de perturbaciones multiplicativas para la pseudoinversa de Moore-Penrose 44
3.3. Perturbaciones multiplicativas para problemas de mínimos cuadrados 48
3.3.1. ¿Por qué el factor $\|A^\dagger\|_2 \|b\|_2/\|x_0\|_2$ es pequeño? 52
3.3.2. El número de condición para perturbaciones multiplicativas del problema de mínimos cuadrados 54
3.3.3. Cotas de perturbaciones multiplicativas para otras soluciones del problema de mínimos cuadrados 55
3.4. Perturbación del problema de mínimos cuadrados en los factores . . . 56
3.5. Algoritmo y análisis de errores 57
3.6. Experimentos Numéricos 63
3.6.1. Matrices de Cauchy 64
3.6.2. Matrices de Vandermonde 66
3.6.3. Matrices Graduadas 69
3.6.4. Experimentos numéricos controlando el residuo 75
3.6.5. Experimentos donde $\|A^\dagger\|_2 \|b\|_2 / \|x_0\|_2$ no es pequeño 76

4. Cálculo de autovalores y autovectores precisos de matrices simétricas graduadas 79
4.1. Matrices simétricas indefinidas y método de pivotaje diagonal 83
4.2. Teoría de perturbaciones de matrices simétricas graduadas 86
4.3. Algoritmos y análisis de errores 91
4.4. Experimentos y ejemplos numérico 99
4.4.1. Experimento numérico 100
4.4.2. Ejemplo numérico 100

5. Conclusiones y trabajos futuros 105

Bibliografía 109

Índice de figuras

3.6.1.Error relativo progresivo $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ frente $\kappa_2(C)$. C matrices de Cauchy aleatorias de orden 100×50. 67

3.6.2.Error relativo progresivo $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ frente n, para matrices de Cauchy de orden $m \times n$ con $m = 50$ y $n = 10{:}2{:}40$ 68

3.6.3.Error relativo progresivo $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ frente $\kappa_2(V)$ para matrices de Vandermonde aleatorias V de tamaños $100 \times 10{:}5{:}60$. 70

3.6.4.Error relativo progresivo $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ frente $\kappa_2(B)$ para matrices graduadas aleatorias $A = S_1 B S_2$ de tamaño 100×40. 73

4.4.1.Error relativo progresivo en los autovalores. Φ como en la ecuación 4.4.3 frente τ para matrices simétricas graduadas aleatorias $A = SBS$ de orden 100. 101

4.4.2.Error relatico progresivo en los autovectores. Ψ como en la ecuación 4.4.4 frente τ para matrices simétricas graduadas aleatorias $A = SBS$ de orden 100. 102

Tabla de Símbolos

Para comodidad del lector presentamos las siguientes tablas de notaciones, que se seguirán a lo largo de toda la tesis.

Símbolo	
$\mathbb{R}, \mathbb{C}$	Conjunto de los números reales, complejos
$\mathbb{R}^n, \mathbb{C}^n$	Conjunto de vectores de n componentes en $\mathbb{R}$, $\mathbb{C}$
$\mathbb{R}^{m\times n}, \mathbb{C}^{m\times n}$	Conjunto de matrices de m filas y n columnas con componentes en $\mathbb{R}$, $\mathbb{C}$
I_n	Matriz identidad de orden n
A^T	Traspuesta de A
$rango(A)$	Rango de A
$\mathcal{R}(A)$	Espacio columna de A
A^{-1}	Inversa de A
$A^\dagger$	Pseudoinversa de Moore-Penrose de A
$A \leq B$	Desigualdades matriciales componente a componente: $a_{ij} \leq b_{ij}\ \forall i, j$
$\lvert A\rvert$	Valor absoluto de A: $\lvert A\rvert_{ij} = \lvert a_{ij}\rvert\ \forall i, j$
$\theta(x, y)$	Ángulo agudo entre los vectores x e y
P_X	Proyección ortogonal sobre el subespacio X
λ_i	i-ésimo autovalor de A
$\widehat{\lambda}_i$	i-ésimo autovalor calculado de A
σ_i	i-ésimo valor singular de A
$\widehat{\sigma}_i$	i-ésimo valor singular calculado de A
$\lVert\cdot\rVert_2$	Norma vectorial 2 o Euclídea: $\lVert x\rVert_2 = (x^T x)^{1/2}$
$\lVert\cdot\rVert_F$	Norma matricial de Frobenius: $\lVert A\rVert_F = \left(\sum_{i=1}^m \sum_{j=1}^n \lvert a_{ij}\rvert^2\right)^{1/2}$
$\lVert\cdot\rVert_2$	Norma matricial 2 o norma espectral: $\lVert A\rVert_2 = \sigma_{\text{máx}}(A)$
$O(\cdot)$	Notación O-grande de Landau
$\mathtt{u}$	Unidad de redondeo
$\kappa(A)$	Número de condición de A: $\kappa(A) = \lVert A\rVert\lVert A^{-1}\rVert$
$relgap_\lambda(A, \lambda_i)$	gap relativo en los autovalores: $relgap_\lambda(A, \lambda_i) = \text{mín}_{j\neq i} \frac{\lvert\lambda_j - \lambda_i\rvert}{\lvert\lambda_i\rvert}$
$relgap_\sigma(A, \sigma_i)$	gap relativo en los valores singulares: $relgap_\sigma(A, \sigma_i) = \text{mín}_{j\neq i} \frac{\lvert\sigma_j - \sigma_i\rvert}{\sigma_i}$

Tabla de Siglas

Siglas	
ALN	Álgebra Lineal Numérica
HRA	Alta Precisión Relativa (High Relative Accuracy)
RRD	Descomposición que revela el rango (Rank revealing decomposition)
EVP	Problema de autovalores (Eigenvalue problem)
SEVP	Problema de autovalores de matrices simétricas (Symmetric eigenvalue problem)
SVD	Descomposición en valores singulares (Singular value decomposition)
LSE	Sistemas de ecuaciones lineales (Linear systems equations)
LSP	Problema de mínimos cuadrados (Least squares problems)
GE	Eliminación Gaussiana (Gaussian elimination)
GECP	Eliminación Gaussiana estándar con pivote completo (GE complete pivoting)
GEPP	Eliminación Gaussiana estándar con pivote parcial (GE partial pivoting)
BP-LDL^T	Factorización simétrica por bloques de Bunch y Parlett con pivote completo
flops	Operaciones en coma flotante (Floating point operations)

Resumen

Esta tesis se enmarca dentro del campo de la *Alta Precisión Relativa (HRA)* en Álgebra Lineal Numérica (ALN). Sus líneas maestras son dos. Por un lado, el diseño y análisis de algoritmos que permitan resolver problemas de Álgebra Lineal con *más precisión de la habitual* para *matrices con estructura*. Y por otro el estudio de la *teoría específica de perturbaciones* necesaria para tratar los problemas que nos ocupan. En nuestra investigación hemos tratado dos:

- La obtención de soluciones precisas del problema de mínimos cuadrados para matrices con estructura (Capítulo 3).
- La obtención de autovalores y autovectores precisos para *matrices simétricas graduadas* (Capítulo 4).

Clásicamente, el Álgebra Lineal Numérica (ALN) es una disciplina de los Métodos Numéricos que desarrolla algoritmos eficientes y estables para

- Resolver sistemas de ecuaciones lineales (LSE): $Ax = b$ con $A \in \mathbb{C}^{n\times n}$ y $b \in \mathbb{C}^n$,
- Resolver problemas de mínimos cuadrados (LSP): $\min_{x\in\mathbb{C}^n} \|Ax - b\|_2$, donde $\|\cdot\|_2$ es la norma euclídea usual, $A \in \mathbb{C}^{m\times n}$ y $b \in \mathbb{C}^m$,
- Calcular autovalores y autovectores de matrices (EVP): $Ax = \lambda x$, $A \in \mathbb{C}^{n\times n}$, $x \in \mathbb{C}^n$ y $\lambda \in \mathbb{C}$, y
- Calcular la Descomposición en Valores Singulares (SVD) de matrices: $A = U\Sigma V^*$ con $A \in \mathbb{C}^{m\times n}$, U y V matrices unitarias y $\Sigma \in \mathbb{C}^{m\times n}$ diagonal con entradas no negativas.

Los cuatro problemas clásicos del Álgebra Lineal Numérica (ALN) siguen siendo objeto de intensa investigación debido a cuatro causas principales relacionadas entre sí:

(a) las aplicaciones actuales requieren cálculos eficientes con matrices cada vez más grandes para las que los algoritmos existentes son muy lentos;

(b) las arquitecturas de los ordenadores evolucionan continuamente, lo que obliga a crear o modificar algoritmos para buscar la máxima eficiencia;

(c) los requerimientos de precisión cada vez son mayores;

(d) aparecen continuamente nuevas clases de matrices estructuradas para las cuales se debe intentar desarrollar algoritmos específicos más eficientes y/o precisos que los existentes.

Hoy en día el *ALN es una disciplina muy amplia* que engloba muchos otros problemas motivados por las aplicaciones. Los problemas clásicos y modernos del ALN están completamente permeados de la idea clave de *aprovechar la estructura concreta de las clases de matrices que aparecen en las aplicaciones en el desarrollo de algoritmos específicos* para dichas matrices, que resuelvan numéricamente los problemas con más rapidez y/o precisión que los algoritmos válidos para matrices generales y que reduzcan la memoria utilizada por el ordenador.

Los mejores algoritmos actualmente existentes en el ALN son *estables en sentido regresivo*, pero esto, a veces, no es suficiente porque se pueden cometer errores inadmisiblemente grandes cuando se aplican a matrices con números de condición muy grandes. El objetivo general de la investigación en algoritmos de alta precisión dentro del ALN es aprovechar la estructura de ciertas clases de matrices para calcular magnitudes con un error mucho menor que el de los algoritmos tradicionales válidos para matrices generales y *esencialmente con el mismo coste computacional* que ellos, esto es, coste $O(n^3)$ *operaciones en coma flotante (flops)* en matrices $n \times n$.

Alcanzar alta precisión (HRA) *sólo es posible para tipos particulares de matrices mediante algoritmos específicos para las mismas* que explotan la estructura del problema para acelerar la velocidad de los cálculos, disminuir los requerimientos de memoria y mejorar la precisión de la solución en comparación con algoritmos estándar.

Nuestra investigación se enmarca en un programa general para encontrar soluciones precisas de los *cuatro problemas clásicos del Álgebra Lineal Numérica*, aprovechando las propiedades de las Descomposiciones que Revelan el Rango (RRD)

usando un esquema básico de algoritmo de *2 pasos* que puede ser usado para resolver diferentes problemas con HRA:

Algoritmo (Algoritmo de *2 pasos* para obtener HRA)

`Paso 1:` Calcular una RRD precisa de $A = XDY$.

`Paso 2:` Aplicar algoritmos, específicos a cada problema tratado, a los factores X, D, Y para obtener la respuesta.

La ventaja de este esquema es su modularidad y que es adaptable a distintos problemas y tipos de matrices. En principio permite obtener HRA para cualquier tipo de matriz para la que se pueda obtener una RRD precisa.

Este esquema de Algoritmo en *2 pasos*, partiendo de una RRD precisa, se aplica ya a la SVD, a Sistemas de Ecuaciones Lineales (LSE), al Problema Simétrico de Autovalores (SEVP). Nosotros lo hemos ampliado al Problema de Mínimos Cuadrados (LSP), Capítulo 3, y al Problema Simétrico de Autovalores (SEVP) para matrices graduadas, Capítulo 4.

En el Capítulo 3 se ha desarrollado una nueva teoría de perturbaciones multiplicativa para la pseudoinversa de Moore-Penrose (Sección 3.2) y para la solución de mínima longitud del problema de Mínimos Cuadrados (Sección 3.3), y se ha implementado un algoritmo para el LSP y se ha realizado su análisis de errores (Sección 3.5), y los correspondientes experimentos numéricos (Sección 3.6).

El resultado que dan los métodos clásicos para el error de la solución de mínima longitud de un problema de Mínimos Cuadrados es:

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \mathtt{u}\, g(m,n)\, \kappa_2^2(A), \tag{0.0.1}$$

donde $g(m,n)$ es una función moderadamente creciente de m y n. Esta cota no garantiza ningún dígito de precisión en la solución calculada si la matriz A está mal condicionada. Desafortunadamente, muchos tipos de matrices estructuradas que aparecen en las aplicaciones están mal condicionadas y los algoritmos convencionales calcularán soluciones del problema de mínimos cuadrados con errores relativos progresivos grandes. La nueva teoría de perturbaciones multiplicativas, unida al análisis de errores del algoritmo que proponemos, muestran que el error progresivo cometido en la solución de mínima longitud (válido para problemas de rango completo, rango

deficiente, o incluso para sistemas lineales infraterminados $m < n$) viene dado por:

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq c\,\mathtt{u}\left[p_y(m,n)\,\kappa_2(Y) + p_x(m,n)\,\kappa_2(X)\,\frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2}\right] + O(\mathtt{u}^2)\,, \quad (0.0.2)$$

donde c es una constante entera pequeña, $p_y(m,n)$ y $p_x(m,n)$ son funciones de m y n moderadamente creciente. El único factor potencialmente grande es $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$, pero un análisis cuidadoso de este factor muestra que es pequeño para la mayoría de vectores b.

En el Capítulo 4 se muestra el trabajo realizado para encontrar soluciones precisas del Problema Simétrico de Autovalores (SEVP) para matrices graduadas. Una matriz simétrica $A = A^T \in \mathbb{R}^{n\times n}$ es llamada graduada si $S^{-1}AS^{-1} \equiv B$ es una matriz bien condicionada para alguna matriz diagonal escalada $S = diag[s_1, \ldots, s_n]$.

En primer lugar se ha demostrado una nueva teoría de perturbaciones estructurada para matrices de la forma $A = SBS$. En la Sección 4.2 se presenta el Teorema 4.2.3 que nos da la sensibilidad del problema a perturbaciones de tipo $\widetilde{A} = S(B + \delta B)S$. Se usa la técnica de transformar perturbaciones aditivas en perturbaciones multiplicativas. Se encuentra que la sensibilidad del problema viene gobernada, entre otros, por un factor nuevo τ_D. La sensibilidad bajo perturbaciones de ese tipo depende del número de condición de los correspondientes factores LDL^T de B y de los elementos de la matriz diagonal S de dos maneras: su orden, después del pivotaje, y el tamaño relativo de los elementos consecutivos en las posiciones de los bloques 2×2 de la matriz D. Este efecto es completamente nuevo y no se ha tenido en cuenta en previos análisis.

El algoritmo que se ha usado en este caso también ha constado de *2 pasos*. Con el fin de calcular una RRD se ha usado la factorización simétrica por bloques $PAP^T = LDL^T$ con estrategia de pivote completo de Bunch y Parlett (factorización BP-LDL^T). Para el segundo paso se ha usado el Algoritmo de Jacobi implícito [30]. El análisis de errores del algoritmo junto con la teoría de perturbaciones multiplicativas muestra que los autovalores y los autovectores se calculan con errores que, a primer orden en la unidad de redondeo $\mathtt{u}$, son

$$\frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq q(n)\,\mathtt{u}\left(\tau\,\Xi_B + \kappa_2(\widehat{L})\right) + O(\mathtt{u}^2)$$
$$\sin\theta(q_i, \widehat{q}_i) \leq q(n)\,\mathtt{u}\left(\tau\,\Xi_B + \kappa_2(\widehat{L})\right)\left(1 + \frac{2}{relgap_{\widehat{\lambda}}(A, \lambda_i)}\right) + O(\mathtt{u}^2)$$

donde Ξ_B es una cantidad pequeña si A está bien escalada, $\widehat{L}$ es el factor L de la factorización BP-LDL^T de A y $\theta(q_i, \widehat{q}_i)$ es el ángulo agudo entre los autovectores

exacto y calculado, respectivamente, q_i y $\widehat{q}_i$. El factor τ controla el escalamiento y es definido como el máximo de tres factores,

$$\tau := max\{1, \tau_L, \tau_D\}, \quad \tau_L := \max_{j<k} \frac{s_k}{s_j} \quad \text{y} \quad \tau_D := \max_{blocks,i} \max \left\{ \frac{s_{i+1}}{s_i}, \frac{s_i}{s_{i+1}} \right\}.$$

El resultado anterior demuestra, en contra de la visión tradicional basada en el caso definido positivo, que no es suficiente que B esté bien condicionada y que los elementos diagonales de la matriz escalada estén ordenados decrecientemente. Si A es una matriz bien escalada en el sentido usual, con los elementos diagonales de S ordenados decrecientemente, entonces $\tau_L \leq 1$, pero el *nuevo* factor τ_D nos dice que esto no es suficiente para obtener alta precisión. El factor τ_D proviene de la presencia de los bloques 2×2 en la factorización de Bunch & Parlett de $A = LDL^T$. Si existe un bloque 2×2 en las posiciones i e $i + 1$ en la matriz diagonal por bloques D y si existe un "salto", ya sea aumentando o disminuyendo en los elementos diagonales de S, s_i y s_{i+1}, habrá un "condicionamiento efectivo"de tamaño τ_D que amplifica la perturbación de entrada de tamaño relativo $\mathtt{u}$. Éste es un nuevo fenómeno.

1 Introducción

Esta tesis se enmarca dentro del campo de la *Alta Precisión Relativa (HRA)* en Álgebra Lineal Numérica (ALN). Sus líneas maestras son dos. Por un lado, el diseño y análisis de algoritmos que permitan resolver problemas de Álgebra Lineal con *más precisión de la habitual* para *matrices con estructura*. Y por otro el estudio de la *teoría específica de perturbaciones* necesaria para tratar los problemas que nos ocupan. Aunque se mencionarán otros problemas relacionados, los que hemos tratado en nuestra investigación son dos:

- La obtención de soluciones precisas del problema de mínimos cuadrados para matrices con estructura (Capítulo 3).
- La obtención de autovalores y autovectores precisos para *matrices simétricas graduadas* (Capítulo 4).

Esta tesis presenta dos características distintivas que no es frecuente encontrar unidas: (a) presta especial atención a las *demostraciones matemáticas de estabilidad y precisión* de los algoritmos considerados –lo que incluye las correspondientes *teorías de perturbaciones* de los problemas tratados; y (b) engloba todos los aspectos del Álgebra Lineal Numérica (ALN): el desarrollo de la teoría necesaria, el desarrollo de nuevos algoritmos, sus análisis de estabilidad y su implementación práctica en el ordenador.

Antes de presentar nuestro trabajo sobre los temas anteriores, daremos una breve panorámica del estado actual del campo en el que se sitúa nuestra investigación.

Clásicamente, el Álgebra Lineal Numérica (ALN) es una disciplina de los Métodos Numéricos que desarrolla algoritmos eficientes y estables para

- Resolver sistemas de ecuaciones lineales (LSE): $Ax = b$ con $A \in \mathbb{C}^{n\times n}$ y $b \in \mathbb{C}^n$,
- Resolver problemas de mínimos cuadrados (LSP): $\min_{x\in\mathbb{C}^n} \|Ax - b\|_2$, donde $\|\cdot\|_2$ es la norma euclídea usual, $A \in \mathbb{C}^{m\times n}$ y $b \in \mathbb{C}^m$,
- Calcular autovalores y autovectores de matrices (EVP): $Ax = \lambda x$, $A \in \mathbb{C}^{n\times n}$, $x \in \mathbb{C}^n$ y $\lambda \in \mathbb{C}$, y
- Calcular la Descomposición en Valores Singulares (SVD) de matrices: $A = U\Sigma V^*$ con $A \in \mathbb{C}^{m\times n}$, U y V matrices unitarias y $\Sigma \in \mathbb{C}^{m\times n}$ diagonal con entradas no negativas.

Estos problemas son *los problemas clásicos del Álgebra Lineal Numérica*. Existen muchos algoritmos numéricos para resolver estos problemas clásicos. Un tratamiento enciclopédido de los mismos puede encontrarse en la referencia fundamental de Golub y Van Loan [39].

Tras algunos antecedentes en el siglo XIX de la mano de prestigiosos matemáticos como Gauss y Jacobi, el comienzo de la investigación sistemática en ALN puede situarse, sin mucha precisión, en la década de 1950 y sus logros desde entonces han sido espectaculares. Cabe destacar en este sentido que en el año 2000 la revista SIAM News de la *Society for Industrial and Applied Mathematics* público el artículo *The Best of the 20th Century: Editors Name Top 10 Algorithms* [15], donde J. Dongarra y F. Sullivan dieron una lista de los que, en su opinión, eran los 10 algoritmos más importantes del siglo XX. De entre ellos, 4 eran algoritmos de ALN (métodos de Krylov, factorizaciones matriciales LU y QR, el algoritmo QR para autovalores y la transformada rápida de Fourier). Podemos decir, por lo tanto, que al menos para los cuatro problemas clásicos mencionados anteriormente existen algoritmos eficientes y estables disponibles en librerías muy prestigiosas como LAPACK [1], ARPACK [50] y MATLAB.

A pesar de ello *los cuatro problemas clásicos del Álgebra Lineal Numérica (ALN)* siguen siendo objeto de intensa investigación debido a cuatro causas principales relacionadas entre sí:

(a) las aplicaciones actuales requieren cálculos eficientes con matrices cada vez más grandes para las que los algoritmos existentes son muy lentos;

(b) las arquitecturas de los ordenadores evolucionan continuamente, lo que obliga a crear o modificar algoritmos para buscar la máxima eficiencia;

(c) los requerimientos de precisión cada vez son mayores;

(d) aparecen continuamente nuevas clases de matrices estructuradas para las cuales se debe intentar desarrollar algoritmos específicos más eficientes y/o precisos que los existentes.

Hoy en día el *ALN es una disciplina muy amplia* que engloba muchos otros problemas aparte de los clásicos. La mayoría de estos nuevos problemas están motivados por las aplicaciones que se están expandiendo desde las áreas clásicas (solución numérica de ecuaciones diferenciales, estadística, optimización, control,...) a nuevas áreas como minería de datos, reconocimiento de patrones, procesamiento de imágenes, física cuántica, etc. Muchas de estas nuevas aplicaciones exigen calcular magnitudes matriciales distintas de las clásicas y, por lo tanto, el desarrollo de nuevos algoritmos.

Los problemas clásicos y modernos del ALN están completamente permeados de la idea clave de *aprovechar la estructura concreta de las clases de matrices que aparecen en las aplicaciones en el desarrollo de algoritmos específicos* para dichas matrices, que resuelvan numéricamente los problemas con más rapidez y/o precisión que los algoritmos válidos para matrices generales y que reduzcan la memoria utilizada por el ordenador. Obviamente estos objetivos pueden no lograrse simultáneamente: los requerimientos de precisión pueden limitar los de rapidez, o los de almacenamiento, etc. Esta idea clave de aprovechar la estructura está en contraste con la investigación desarrollada en las primeras décadas del ALN ($\sim 1950 - 1980$), en las que la mayor parte de los avances se centraron en algoritmos generales válidos para matrices cualesquiera (eliminación Gaussiana, factorización QR para problemas de mínimos cuadrados, algoritmo QR para autovalores, bidiagonalización de Golub y Kahan para valores singulares). Las referencias [58, 59], junto con la bibliografía en ella citada, constituye un buen punto de partida para comprender la importancia de las matrices estructuradas en el ALN moderna.

Los mejores algoritmos actualmente existentes pueden cometer errores inadmisiblemente grandes cuando se aplican a ciertas matrices que son importantes en la práctica, por lo tanto es necesario recordar cómo son los errores cometidos por estos algoritmos tradicionales del ALN y por qué es necesario mejorarlos. Para evitar entrar en una discusión muy extensa nos centraremos en los errores en los algoritmos directos para calcular autovalores de matrices simétricas (SEVP) y para resolver sistemas de ecuaciones lineales (SEL), sin embargo dificultades similares aparecen en todos los problemas del ALN.

Los mejores algoritmos del ALN son *estables en sentido regresivo*. Por simplicidad, explicaremos este concepto considerando el caso particular de los algoritmos para calcular los autovalores, $\lambda_1 \geq \cdots \geq \lambda_n$, de una matriz simétrica $A = A^T \in \mathbb{R}^{n\times n}$. Diremos que un algoritmo para el cálculo de autovalores es *regresivamente estable* si, para cada matriz A, los autovalores de A calculados mediante el algoritmo, son los autovalores exactos de una matriz ligeramente perturbada

$$A + E \tag{1.0.1}$$

(esto es, la matriz E es pequeña, por ejemplo en norma, comparada con la matriz A). Si para calcular los autovalores de una matriz simétrica real A empleamos un método regresivamente estable, los autovalores calculados por dicho método $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n$, serán los autovalores exactos de una matriz perturbada $A + E$ también simétrica, donde la matriz E acumula, por decirlo de alguna manera, todos los errores debidos al redondeo en las operaciones aritméticas que requiere el proceso del cálculo de autovalores

Lo primero que se debe notar sobre un algoritmo estable en sentido regresivo es que la ecuación (1.0.1), no nos da el error cometido en los autovalores calculados. Para acotar dicho error necesitamos un resultado de perturbación matricial que relacione los autovalores de A (los exactos) con los de $A + E$ (los calculados). En el caso que nos ocupa este es el teorema de perturbación de Weyl [17], que implica que

$$|\widehat{\lambda}_i - \lambda_i| \leq \|E\|_2 = O(\mathtt{u})\|A\|_2 \quad \forall i, \tag{1.0.2}$$

donde $\mathtt{u}$ es la unidad de redondeo del ordenador ($\mathtt{u} \approx 10^{-16}$ si se calcula en doble precisión).

Los algoritmos disponibles en LAPACK [1] para el problema espectral simétrico (QR, Divide y Conquista, Multiple Relatively Robust Representations (MRRR) y el algoritmo clásico de Jacobi [17]) son estables en sentido regresivo. También lo es el algoritmo utilizado por MATLAB (QR). Sin embargo queremos dejar claro que no todos los algoritmos del ALN son estables en sentido regresivo y que aquellos que lo son se consideran verdaderas joyas del Análisis Numérico y muy satisfactorios desde el punto de vista de los errores cometidos. La razón es que, como en la práctica las entradas de una matriz suelen venir afectadas de errores, la precisión alcanzada por dichos algoritmos simplemente refleja esta incertidumbre. De (1.0.2), teniendo en cuenta que para una matriz simétrica su norma coincide con el módulo del autovalor $\lambda_{\max}$ de módulo máximo de la matriz, obtenemos la siguiente cota de error relativo

para los autovalores,

$$\frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} = O(\mathtt{u})\frac{|\lambda_{\max}|}{|\lambda_i|} \quad \forall i, \tag{1.0.3}$$

que será muy grande si $\frac{|\lambda_{\max}|}{|\lambda_i|} \gtrsim 10^{16}$. Nótese que para el autovalor de módulo mínimo $\lambda_{\min}$, el cociente

$$\frac{|\lambda_{\max}|}{|\lambda_{\min}|} = \|A\|_2 \, \|A^{-1}\|_2 \equiv \kappa_2(A) \tag{1.0.4}$$

es precisamente el número de condición de la matriz [39] en la norma espectral. Por lo tanto, los errores cometidos pueden ser muy grandes para matrices mal condicionadas.

Algo semejante ocurre en otro caso muy conocido, la solución de sistemas de ecuaciones lineales. Consideremos $A \in \mathbb{C}^{n\times n}$ y no singular, y sea $\widehat{x}$ la solución del sistema de ecuaciones $Ax = b$ calculada por medio de la factorización LU utilizando eliminación Gaussiana con pivote parcial. El error regresivo está dado por [43, Teorema 9.4]:

$$(A + E)\widehat{x} = b, \quad \text{con} \quad |E| = O(\mathtt{u})\,|\widehat{L}||\widehat{U}|, \tag{1.0.5}$$

donde $\widehat{L}$ y $\widehat{U}$ son los factores calculados por la factorización LU. Ahora bien, situándonos en un caso favorable, si $\widehat{L}$ y $\widehat{U}$ son factores no negativos entonces $|\widehat{L}||\widehat{U}| = |\widehat{L}\widehat{U}|$, por la tanto la cota dada en la ecuación (1.0.5) se convierte en:

$$(A + E)\widehat{x} = b, \quad \text{con} \quad |E| = O(\mathtt{u})\,|A|. \tag{1.0.6}$$

La teoría de perturbaciones de LSE es bien conocida, las soluciones de $Ax = b$ y $(A + E)\widetilde{x} = b$ están relacionadas de la forma

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|E\|}{1 - \|A^{-1}\|\|E\|}. \tag{1.0.7}$$

Usando el mismo razonamiento que en la obtención de la ecuación (1.0.4), concluimos que el error relativo de la solución de sistemas de ecuaciones lineales vía la factorización LU obedece:

$$\frac{\|\widehat{x} - x\|}{\|x\|} = O(\mathtt{u})\,\kappa(A). \tag{1.0.8}$$

La cota dada en la ecuación (1.0.8) no garantiza un sólo dígito de precisión si $\kappa(A) \gtrsim 1/\mathtt{u}$, es decir, si A está mal condicionada con respecto a la inversa de la

unidad de redondeo. Desafortunadamente, muchas de las matrices estructuradas que aparecen en las aplicaciones están muy mal condicionadas. Dos ejemplos clásicos son las matrices de Cauchy y de Vandermonde [43, Capítulos 22 y 28].

El objetivo general de la investigación en algoritmos de alta precisión dentro del ALN es aprovechar la estructura de ciertas clases de matrices para calcular magnitudes con *error mucho menor*[1] que el de los algoritmos tradicionales válidos para matrices generales y *esencialmente con el mismo coste computacional* que ellos, esto es, coste $O(n^3)$ *operaciones en coma flotante (flops)* en matrices $n \times n$. Debe quedar claro que los algoritmos de alta precisión son, en general, más lentos que los tradicionales, pero no mucho más lentos. Ello se refleja en que la dependencia del coste en la dimensión es del mismo tipo. Esta limitación en el coste computacional implica que *el uso de programas de precisión variable o cálculo simbólico está prohibido* en la investigación en algoritmos de alta precisión por su extrema lentitud y por la dependencia del coste computacional del condicionamiento del problema.

Volviendo al caso particular del cálculo de los autovalores de matrices simétricas explicado anteriormente, la meta ideal de un algoritmo de alta precisión es calcular los autovalores con errores

$$\frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} = O(\mathtt{u}) \quad \forall i, \tag{1.0.9}$$

en vez de con errores (1.0.3).

De la misma manera, para el caso de sistemas de ecuaciones lineales el objeto de la alta precisión será lograr que la solución calculada $\widehat{x}$ satisfaga que:

$$\frac{\|\widehat{x} - x\|}{\|x\|} = O(\mathtt{u}). \tag{1.0.10}$$

Cuando los autovalores calculados por un algoritmo satisfacen una cota de error como la que aparece en (1.0.9) se dice que el algoritmo alcanza *alta precisión relativa (HRA)*. La diferencia fundamental respecto a los algoritmos tradicionales es que los errores en (1.0.9) y (1.0.10) son siempre pequeños independientemente de la magnitud del número de condición tradicional de la matriz, mientras que los que aparecen en (1.0.3) y (1.0.8), a veces son pequeños (para matrices bien condicionadas) y a veces no (para matrices mal condicionadas).

[1] El significado riguroso de la frase *error mucho menor* depende del problema particular que se considere, así como de los requerimientos de precisión impuestos por la aplicación o el usuario. En sentido general puede decirse que su significado matemático es calcular con un error gobernado por un número de condición mucho menor que el número de condición tradicional para el problema que se esté resolviendo.

Alcanzar alta precisión (HRA) *sólo es posible para tipos particulares de matrices mediante algoritmos específicos para las mismas*. Siguiendo con el caso particular de los autovalores de matrices simétricas (SEVP) ello es fácil de entender intuitivamente: si una matriz no tiene ninguna estructura adicional, las $(n^2 + n)/2$ entradas de su parte triangular inferior son parámetros independientes, no hay razón para esperar más que la estabilidad regresiva tradicional explicada en (1.0.1). Sin embargo, si una matriz depende de unos ciertos parámetros, existe la posibilidad de desarrollar algoritmos que sean *estables en sentido regresivo respecto de los parámetros que definen la matriz o, al menos, que produzcan errores relativos de magnitud como si lo fueran*. Ello restringe enormemente el conjunto de perturbaciones posibles asociadas a los errores regresivos y puede disminuir drásticamente el número de condición respecto de este conjunto restringido de perturbaciones y, por consiguiente, los errores cometidos.

Matrices con estructuras particulares aparecen frecuentemente en la teoría y aplicaciones [58, 59]. Como consecuencia, el diseño y análisis de algoritmos especiales que involucran matrices estructuradas es un área clásica del Álgebra Lineal Numérica que atrae la atención de muchos investigadores. Algoritmos especiales para resolver sistemas de ecuaciones estructurados o problemas estructurados de autovalores están incluidos en muchas referencias estándar, por ejemplo [17, 39, 43, 49, 73].

El objetivo al considerar algoritmos especiales es explotar la estructura del problema para acelerar la velocidad de los cálculos, disminuir los requerimientos de memoria y mejorar la precisión de la solución en comparación con algoritmos estándar.

Las primeras matrices consideradas asociadas al problema de autovalores fueron las matrices tridiagonales y bidiagolanes. El primer análisis de errores de un algoritmo de alta precisión relativa se debe a Kahan [47], un trabajo que tendría su continuación en Demmel y Kahan [21], donde se propone un algoritmo para calcular con alta precisión relativa tanto valores singulares de matrices bidiagonales como autovalores de ciertas matrices simétricas tridiagonales. Un algoritmo alternativo es el algoritmo dqds de Fernando y Parlett [38], que también calcula con alta precisión relativa los valores singulares de matrices bidiagonales, y lo hace con un esfuerzo computacional comparable al de algoritmos de carácter generalgeneral tipo QR. En cuanto al problema de autovalores para matrices tridiagonales, los algoritmos de alta precisión relativa más usados son los propuestos por Dhillon y Parlett [26, 27].

Más tarde fueron consideradas las matrices hermíticas densas; diversos autores diseñaron diferentes algoritmos. En primer lugar, Barlow y Demmel presentan en [3] un algoritmo de alta precisión relativa para autovalores de matrices escaladas diagonal-

mente dominantes. Basándose en ello, Demmel y Veselić obtienen en [25] uno de los resultados más nítidos y llamativos de este área, al demostrar que el algoritmo clásico de Jacobi (con un criterio de parada adecuado) calcula con alta precisión relativa los autovalores de matrices simétricas definidas positivas, algo que ya habían conjeturado, sin demostrarlo, Hari y Veselić en [71]. De hecho, Demmel y Veselić muestran que el algoritmo de Jacobi es óptimo para el caso de las matrices definidas positivas, en el sentido de que los errores introducidos por el algoritmo son del mismo orden que los errores inherentes al proceso de almacenar la matriz en el ordenador. En particular, el algoritmo de Jacobi *es más preciso que cualquier algoritmo, como QR, bisección, o divide y conquista, que inicialmente tridiagonalice la matriz*. Veselić y Slapničar, por su parte, extienden en [66, 70, 72] parte de los resultados de [25] al caso de matrices hermíticas indefinidas, proponiendo un algoritmo de tipo Jacobi que, además de las habituales rotaciones trigonométricas de Jacobi, utiliza rotaciones hiperbólicas. Hay que destacar que en este caso no se alcanza la alta precisión relativa para cualquier tipo de matriz hermítica, y además el algoritmo no es capaz de proporcionar alta precisión relativa para todas las matrices que, de acuerdo con la teoría, lo admiten. Esto se logró con posterioridad con dos nuevos algoritmos. En [33] Dopico, Molera y Moro presentaron un algoritmo, basado en una SVD previa, que obtenía HRA para cualquier matriz para la que se pudiera obtener una RRD precisa. Y en 2009, Dopico, Koev y Molera [30] presentaron un algoritmo, de Jacobi implícito, que mantenía la simetría del problema y que también alcanzaba HRA para cualquier matriz para la que se pudiera obtener una RRD precisa. Entre otras referencias, para los problemas de la SVD (Descomposición en Valores Singulares) y el SEVP (Problema Simétrico de autovalores) ver también [35, 36, 48, 75].

Algoritmos especiales para resolver sistemas de ecuaciones estructurados más precisos que los métodos estándar aparecieron desde los inicios del *Álgebra Lineal Numérica*, por ejemplo en [6], Björck y Pereyra, encuentran las soluciones cuando la matriz de coeficientes es una matriz de Vandermonde, en [32] Dopico y Molera, calculan soluciones exactas para las matrices estructuradas a las que es posible calcular una RRD precisa.

Algoritmos para calcular soluciones del problema de mínimos cuadrados $\min_x \|b - Ax\|_2$, donde $A \in \mathbb{R}^{m\times n}$ es una matriz estructurada, con más precisión que la conseguida por los algoritmos estándar con el mismo coste computacional, es decir, $O(n^2m)$ flops, han recibido atención cuando la matriz A es diagonalmente escalada [4, 16, 44, 64]. Para otras clases de matrices estructuradas en [53] los autores resuelven el problema para matrices totalmente positivas.

En [3, 18–25, 28, 62, 75] encontramos soluciones precisas del problema de autovalores y del problema de valores singulares, cuando la matriz tiene una estructura particular. Pero en 1999 Demmel, Gu, Eisenstat, Slapničar, Veselić y Drmač [20] realizaron una contribución fundamental en el campo de la alta precisión relativa (HRA) al presentar un punto de vista unificado que permite tratar muchos de los casos anteriores. La conclusión fundamental en [20] es que los valores singulares de una matriz A se pueden calcular con alta precisión relativa si y sólo si la matriz se puede factorizar numéricamente con alta precisión relativa en la forma $A = XDY$ con D diagonal y X e Y bien condicionadas. Observemos que esto significa que si A está mal condicionada, entonces el factor diagonal también lo es. Esto es lo que en [20] se define como una Descomposición que Revela el Rango (*Rank-Revealing-Decomposition, (RRD)*) (ver la Definición 2.4.1 y también [43, Sección 9.12]). Obsérvese que la SVD es un caso particular de RRD (y que en ese caso el resultado de [20] es trivial).

La propiedad fundamental por la cual una RRD es muy útil en cálculos con alta precisión relativa, es que sus factores determinan de manera precisa la SVD de la matriz original, es decir, pequeñas perturbaciones relativas componente a componente en los elementos de la matriz D y pequeñas perturbaciones relativas en norma en X e Y, producen pequeñas perturbaciones en los valores singulares de A y pequeñas perturbaciones en los vectores singulares con respecto al gap relativo de los valores singulares [20]. Además, muestran que la única teoría de perturbaciones que se requiere para tal unificación es la ya desarrollada por Eisenstat e Ipsen en [37] y por Li en [51, 52]. En esta misma línea, Demmel y Koev presentan en [22] condiciones necesarias y suficientes sobre matrices estructuradas para que se pueda calcular con alta precisión relativa su descomposición en valores singulares.

Para resolver el problema de autovalores con alta precisión relativa de matrices simétricas para las cuales es posible calcular una RRD, se han propuesto diversos algoritmos en el pasado. Estos algoritmos están basados en el algoritmo de Jacobi "one-sided" [17, Sección 5.4.3] (ver también la Sección 2.5). Los algoritmos en [25, 34, 55] proporcionan HRA para matrices simétricas definidas positivas; y el algoritmo de Jacobi Implícito [30] lo hace para matrices simétricas generales . En ambos casos el uso de un adecuado criterio de parada es crucial.

Calcular una RRD precisa $A = XDY$ no es siempre sencillo. Para casi cualquier matriz $A \in \mathbb{C}^{m\times n}$ una RRD (potencialmente imprecisa) puede calcularse aplicando eliminación Gaussiana estándar con pivote completo (GECP) para obtener, excepto permutaciones, una factorización LDU, donde $X = L \in \mathbb{C}^{m\times r}$ es trapezoidal inferior

unitaria (notación de [43, p. 355]), $D \in \mathbb{C}^{r\times r}$ es diagonal y no singular e $Y = U \in \mathbb{R}^n$ es trapezoidal superior unitaria [20, 43]. Otra opción es usar el algoritmo QR de Householder con pivote por columnas y tomar $X = Q$, $D = diag(R)$ e $Y = D^{-1}R$, excepto permutaciones. Muy raramente GECP o QR con pivote por columnas fallan para producir factores X e Y bien condicionados. En [40, 57, 60] podemos encontrar otras estrategias de pivotaje que garantizan factores bien condicionados. Sin embargo ni GECP estándar ni QR con pivote por columnas son precisas para matrices mal condicionadas y, actualmente, RRDs con precisión garantizada pueden calcularse sólo para clases particulares de matrices estructuradas a través de GECP que explotan cuidadosamente la estructura para obtener factores precisos, y en el caso de matrices graduadas (esto es, matrices de la forma S_1BS_2 con B bien condicionada, S_1 y S_2 diagonales), utilizar factorización QR de Householder con pivote completo [42].

El cálculo de RRDs precisas es posible para muchas clases de matrices simétricas: matrices diagonalmente dominantes [3], matrices definidas positivas diagonalmente escaladas y bien condicionadas [25], algunas matrices indefinidas diagonalmente escaladas y bien condicionadas [65], M-matrices diagonalmente dominantes débiles [23], matrices de Cauchy, matrices de Cauchy diagonalmente escaladas, matrices de Vandermonde, matrices totalmente no negativas [28], "total signed compound matrices", matrices escaladas diagonalmente totalmente unimodulares [62], matrices diagonalmente dominantes bien parametrizadas [75]. Gracias a la intensa investigación desarrollada en los últimos 20 años sobre cálculo de SVDs con alta precisión relativa, existen algoritmos para calcular RRDs precisas de muchas clases de matrices $m \times n$ estructuradas en $O(m\,n^2)$ operaciones. Estas clases incluyen incluso algunas de las existentes inicialmente sólo para matrices simétricas: matrices de Cauchy, matrices de Cauchy diagonalmente escaladas, matrices de Vandermonde y algunas matrices "related unit-displacement-rank" [18]; matrices graduadas [20, 42]; matrices acíclicas (incluyen matrices bidiagonales), "total signed compound matrices", matrices diagonalmente escaladas totalmente unimodulares [20]; M-matrices diagonalmente dominantes [23, 63]; matrices polinomiales de Vandermonde que involucran polinomios ortonormales [24]; matrices diagonalmente dominantes [29, 75].

Se puede afirmar que nuestra investigación se enmarca en un programa general para encontrar soluciones precisas de los *cuatro problemas clásicos del Álgebra Lineal Numérica*, aprovechando las propiedades de la RRD (revisamos el significado de RRD "precisa" en la Definición 2.4.2) dadas en [20]. En [20] se introdujo el esquema básico de algoritmo de *2 pasos* que puede ser usado para resolver diferentes problemas con HRA:

Algoritmo 1.0.1 (Algoritmo de *2 pasos* para obtener HRA)

`Paso 1:` Calcular una RRD precisa de $A = XDY$.

`Paso 2:` Aplicar algoritmos, específicos a cada problema tratado, a los factores X, D, Y para obtener la respuesta.

Por ejemplo, en [20], para resolver el problema de la SVD se aplicó, una combinación de descomposición QR y Jacobi "one-sided"(Algoritmo 2.5.4).

La ventaja de este esquema es su modularidad y que es adaptable a distintos problemas y tipos de matrices. En principio permite obtener HRA para cualquier tipo de matriz para la que se pueda obtener una RRD precisa. Como la clase de matrices para las que esto se puede hacer es amplia (véase más arriba), y además susceptible de aumentar con nuevas clases de matrices estructuradas, las posibilidades prácticas son grandes.

En concreto este esquema (Algoritmo en *2 pasos*, partiendo de una RRD precisa) se aplica ya, además de a la SVD [20], a Sistemas de Ecuaciones Lineales (LSE) [32], al Problema Simétrico de autovalores (SEVP) [13, 30] y al Problema de Mínimos Cuadrados (LSP) [11, 12].

En esta tesis hemos trabajado en el Problema de Mínimos Cuadrados (LSP), Capítulo 3, y en el Problema Simétrico de autovalores (SEVP) para matrices simétricas graduadas, Capítulo 4.

En [32] Dopico y Molera trataron el problema de calcular de manera precisa la solución de sistemas lineales de ecuaciones usando el esquema anterior. Nosotros hemos realizado una extensión no trivial de este trabajo, con el fin, de resolver de manera precisa el problema de mínimos cuadrados (Capítulo 3). Esto ha requerido el desarrollo de una nueva teoría de perturbaciones multiplicativa de la pseudoinversa de Moore-Penrose (Sección 3.2) y para la solución de mínima longitud del problema de Mínimos Cuadrados (Sección 3.3), así como la de implementar de forma específica un algoritmo para el LSP y su análisis de errores (Sección 3.5), y experimentos numéricos (Sección 3.6).

En relación con el problema de encontrar soluciones precisas de los sistemas de ecuaciones lineales $Ax = b$, presentamos a continuación el algoritmo, siguiendo la estructura de obtener la solución en dos pasos

Algoritmo 1.0.2 (Algoritmo de *2 pasos* para obtener HRA para LSE)

`Paso 1:` Calcular una RRD precisa de $A = XDY$.

`Paso 2:` Resolver $XDYx = b$ mediante tres pasos:
Resolver $X\,s = b$.
Resolver $Dw = s$ mediante $w_i = s_i/d_i$, $i = 1 : r$.
Resolver $Yx = w$.

El procedimiento descrito anteriormente calcula soluciones precisas, incluso para matrices A mal condicionadas, puesto que, cada entrada de w se calcula con error relativo menor que $\mathtt{u}$, es decir, el sistema lineal mal condicionado $D\,w = s$ se resuelve de manera precisa, y además $Xs = b$ e $Yx = w$ también se resuelven de manera precisa, ya que X e Y son factores bien condicionados. En [32, Teorema 4.2] se demostró que el error relativo para la solución calculada $\widehat{x}$ por el algoritmo descrito anteriormente es

$$\frac{\|\widehat{x} - x\|}{\|x\|} \leq \mathtt{u}\, f(n)\ \text{máx}\{\kappa(X), \kappa(Y)\}\frac{\|A^{-1}\|\,\|b\|}{\|x\|}, \tag{1.0.11}$$

donde $f(n)$ es una función moderadamente creciente de n. Observemos que el único factor potencialmente grande que aparece en la ecuación (1.0.11) es $\|A^{-1}\|\,\|b\|/\|x\|$, ya que X e Y son los factores bien condicionados de una RRD. El análisis cuidadoso de este factor muestra [14, 32] que, aunque puede llegar a valer $\kappa(A)$, es pequeño para la mayoría de vectores b, incluso si A está mal condicionada. Este hecho lo explicaremos cuidadosamente en la Sección 3.3.1.

En el Capítulo 3 se recoge nuestro trabajo para tratar problemas estructurados de mínimos cuadrados $\text{mín}_x \|b - Ax\|_2$, donde $A \in \mathbb{R}^{m\times n}$ y $b \in \mathbb{R}^m$, con más precisión que la conseguida con algoritmos convencionales. Generalmente el problema de mínimos cuadrados $\text{mín}_x \|b - Ax\|_2$ con rango completo por columnas se soluciona por medio de la factorización QR o de la SVD, ambos métodos son regresivamente estables, es decir, la solución calculada por cualquiera de ellos, $\widehat{x}_0$, es la solución exacta del problema de mínimos cuadrados $\text{mín}_x \|(b + \Delta b) - (A + \Delta A)x\|_2$, donde $\|\Delta b\|_2 \leq c\,\mathtt{u}\,mn\,\|b\|_2$, $\|\Delta A\|_2 \leq c\,\mathtt{u}\,m\,n^{3/2}\,\|A\|_2$, $\mathtt{u}$ es la unidad de redondeo del ordenador y c denota una pequeña constante entera [43, Teorema 20.3]. Este resultado fuerte de errores regresivos junto con la teoría de perturbaciones clásica en norma del problema de mínimos cuadrados ([74, Teorema 5.1], [5, Teorema 1.4.6]), implican la siguiente

cota de error progresiva en la solución calculada $\widehat{x}_0$ con respecto a la solución exacta x_0

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \mathtt{u}\, g(m,n)\, \kappa_2(A)^2, \tag{1.0.12}$$

donde $g(m,n)$ es una función moderadamente creciente de m y n. Esta cota no garantiza ningún dígito de precisión en la solución calculada si la matriz A está mal condicionada. Desafortunadamente, muchos tipos de matrices estructuradas que aparecen en las aplicaciones están mal condicionadas y los algoritmos convencionales calcularán soluciones del problema de mínimos cuadrados con errores relativos progresivos grandes. Nuestro objetivo ha sido proponer un algoritmo para resolver problemas estructurados de mínimos cuadrados con alta precisión relativa, esto lo obtendremos a través de la RRD y el algoritmo constará de los dos pasos fundamentales tal y como mencionamos anteriormente:

Algoritmo 1.0.3 (Algoritmo de *2 pasos* para obtener HRA)

`Paso 1:` Calcular una RRD precisa de $A = XDY$.

`Paso 2:` Calcular la solución de mínima longitud de $\text{mín}_x \|XDYx - b\|_2$ mediante tres pasos:

Calcular la solución única s de $\text{mín}_x \|b - Xx\|_2$ vía la factorización QR de Householder.

Calcular la solución w del sistema lineal $Dw = s$ como $w_i = s_i/d_i$, $i = 1 : r$.

Calcular la solución de mínima longitud x_0 del sistema lineal infraterminado $Yx = w$ utilizando el método Q [43, Capítulo 21].

Al estar los factores X e Y bien condicionados, los pasos primero y tercero $\text{mín}_x \|b - Xx\|_2$ e $Yx = x_2$ son resueltos de manera precisa, el único factor mal condicionado es la D, pero el paso donde se utiliza este factor se hace de manera eficiente garantizando soluciones precisas, incluso si el factor D está mal condicionado.

La nueva teoría multiplicativa de la pseudoinversa de Moore-Penrose y para la solución de mínima longitud del problema de Mínimos Cuadrados, unidos al análisis de errores del algoritmo muestran que el error progresivo cometido en la solución de mínima longitud (para problemas de rango completo, rango deficiente, o incluso

para sistemas lineales infraterminados $m < n$) viene dado por:

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq c\,\mathtt{u}\left[p_y(m,n)\,\kappa_2(Y) + p_x(m,n)\,\kappa_2(X)\,\frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2}\right] + O(\mathtt{u}^2)\,, \quad (1.0.13)$$

donde c es una constante pequeña entera[2], $p(m,n)$ es una función de m y n moderadamente creciente, $p_y(m,n) := (p(m,n) + nr^{3/2})$ y $p_x(m,n) := (p(m,n) + mr^{3/2})$; y no con errores (1.0.12). El único factor potencialmente grande es $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$. Al igual que en el caso de LSE mostrado arriba, el análisis cuidadoso de este factor muestra que es pequeño para la mayoría de vectores b (Sección 3.3.1).

En el Capítulo 4 se muestra el trabajo realizado en esta tesis para encontrar soluciones precisas del Problema Simétrico de Autovalores (SEVP) para matrices graduadas. Una matriz simétrica $A = A^T \in \mathbb{R}^{n\times n}$ es llamada graduada si $S^{-1}AS^{-1} \equiv B$ es una matriz bien condicionada para alguna matriz diagonal escalada $S = diag[s_1, \ldots, s_n]$.

En primer lugar se ha demostrado una nueva teoría de perturbaciones estructurada para matrices de la forma $A = SBS$. En la Sección 4.2 se presenta el Teorema 4.2.3 que nos da la sensibilidad del problema a perturbaciones de tipo $\widetilde{A} = S(B + \delta B)S$. Se usa la técnica de transformar perturbaciones aditivas en perturbaciones multiplicativas. Se encuentra que la sensibilidad del problema viene gobernada, entre otros, por un factor nuevo τ_D (ver (4.0.4)). La sensibilidad bajo perturbaciones de ese tipo depende del número de condición de los correspondientes factores $L_B D_B L_B^T$ de B y de los elementos de la matriz diagonal S de dos maneras: su orden, después del pivotaje, y el tamaño relativo de los elementos consecutivos en las posiciones de los bloques 2×2 de la matriz D. Este efecto es completamente nuevo y no se ha tenido en cuenta en previos análisis.

En 2009, Dopico, Koev y Molera [30] presentaron el algoritmo de Jacobi implícito, basándose en los *2 pasos* del Algoritmo 1.0.1 (Ver Algoritmo 2.5.6), y demostraron que se podía calcular con HRA el SEVP para cualquier matriz simétrica para la que se pudiera obtener una RRD precisa. Nosotros hemos analizado el SEVP para un caso concreto de matrices estructuradas simétricas, las matrices graduadas.

El algoritmo que se ha usado en este caso (Algoritmo 4.3.1) ha constado de *2 pasos* como el Algoritmo 1.0.1. Con el fin de calcular una RRD, en este caso se ha usado la factorización simétrica por bloques $PAP^T = LDL^T$ con estrategia de pivote completo de Bunch y Parlett, BP-LDL^T [1, 39, 43]: L es triangular inferior

[2]En esta tesis denotaremos por *c todas las diferentes* constantes enteras pequeñas que van apareciendo.

con unos en la diagonal y en la práctica está bien condicionada, y D es diagonal por bloques, con bloques de dimensión 1 ó 2. Los bloques 2×2 diagonales son matrices indefinidas simétricas, y los correspondientes bloques diagonales de L son matrices identidad de orden 2. Este método también es conocido como *método de pivotaje diagonal* y lo describiremos en la Sección 4.1. Para el segundo paso se ha usado el Algoritmo 2.5.6 (de Jacobi implícito). El algoritmo y su análisis de errores se presentan en la Sección 4.3. Se demuestra que nuestro algoritmo introduce pequeñas perturbaciones multiplicativas regresivas de la matriz, y después se utiliza el hecho bien conocido que estas pequeñas perturbaciones multiplicativas producen pequeñas perturbaciones relativas en los autovalores y autovectores, Teoremas 2.2.2 y 2.2.3. En definitiva el algoritmo propuesto muestra la sensibilidad del problema SEVP para matrices simétricas graduadas $A = SBS$. Los autovalores y los autovectores se calculan con los errores dados en el Corolario 4.3.6, a primer orden en la unidad de redondeo $\mathtt{u}$,

$$\frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq q(n)\,\mathtt{u}\left(\tau\ \Xi_B + \kappa_2(\widehat{L})\right) + O(\mathtt{u}^2) \tag{1.0.14}$$

$$\sin\theta(q_i, \widehat{q}_i) \leq q(n)\,\mathtt{u}\left(\tau\ \Xi_B + \kappa_2(\widehat{L})\right)\left(1 + \frac{2}{relgap_{\widehat{\lambda}}(A, \lambda_i)}\right) + O(\mathtt{u}^2) \tag{1.0.15}$$

donde Ξ_B es una pequeña si A está bien escalada (veáse el Corolario 4.3.6 para su definición), $\widehat{L}$ es el factor L de la factorización BP-LDL^T de A y $\theta(q_i, \widehat{q}_i)$ es el ángulo agudo entre los autovectores exacto y calculado, respectivamente, q_i y $\widehat{q}_i$. El factor τ controla el escalamiento y es definido como el máximo de tres factores,

$$\tau = max\{1, \tau_L, \tau_D\}, \quad \tau_L = \max_{j<k} \frac{s_k}{s_j} \quad \text{y} \quad \tau_D = \max_{blocks,i} \max\left\{\frac{s_{i+1}}{s_i}, \frac{s_i}{s_{i+1}}\right\}. \tag{1.0.16}$$

Previamente no existían resultados sobre las condiciones que $S = diag[s_1, \ldots, s_n]$ y B tienen que satisfacer para obtener soluciones precisas del problema de autovalores de matrices simétricas generales $A = SBS$. El resultado anterior demuestra, en contra de la visión tradicional basada en el caso definido positivo, que no es suficiente que B esté bien condicionada y que los elementos diagonales de la matriz escalada estén ordenados decrecientemente. Si A es una matriz bien escalada en el sentido usual, con los elementos diagonales de S ordenados decrecientemente, entonces $\tau_L \leq 1$, pero el *nuevo* factor τ_D nos dice que esto no es suficiente para obtener alta precisión. El factor τ_D proviene de la presencia de los bloques 2×2 en la factorización de Bunch & Parlett de $A = LDL^T$ (ver Sección 4.1). Si existe un bloque 2×2 en las

posiciones i e $i+1$ en la matriz diagonal por bloques D y si existe un "salto", ya sea aumentando o disminuyendo en los elementos diagonales de S, s_i y s_{i+1}, habrá un "condicionamiento efectivo"de tamaño τ_D que amplifica la perturbación de entrada de tamaño relativo $\mathtt{u}$. Observemos que esto es un nuevo fenómeno, que no aparece por ejemplo en las perturbaciones de valores singulares de matrices graduadas [20].

La precisión de este algoritmo es comprobada por medio de experimentos numéricos en la Sección 4.4.

Para finalizar esta Sección es importante mencionar los artículos a que ha dado lugar la investigación conducente a esta tesis: el contenido del Capítulo 3 corresponde a los artículos [11, 12]. El artículo [11], ha sido aceptado recientemente por la revista *SIAM Journal on Matrix Analysis*, y el artículo [12] se someterá en breve a la revista *Linear Algebra and its Applications*. Así mismo, el contenido de los artículos [11, 12] se han presentado en los siguientes Congresos Internacionales: *Householder Symposium XVIII, 7th International Congress on Industrial and Applied Mathematics, The 17th International Linear Algebra Society Conference* y *2012 SIAM Conference on Applied Linear Algebra*.

El contenido del Capítulo 4 corresponde al artículo [13] que ha sido enviado a la revista *Electronic Transactions on Numerical Analysis* y presentado en los Congresos Internacionales: *XVII Congreso Colombiano de Matemáticas* y *Álgebra Lineal, Análisis Matricial y Aplicaciones - ALAMA 2012*.

2

Descomposiciones que revelan el rango y alta precisión relativa: Preliminares y resultados previos

En este capítulo vamos a presentar los resultados y definiciones necesarias en los que está basado nuestro trabajo, que se expondrá en los capítulos siguientes. Empezaremos con una sección dedicada a la notación que va a ser empleada a lo largo de la memoria (ver también el Tabla de Símbolos y Siglas). A continuación expondremos resultados conocidos con anterioridad sobre teoría de perturbaciones que nos sirven de base y de referencia. Haremos especial hincapié en el concepto de perturbaciones multiplicativas y la factorización de matrices conocida en la literatura como RRD (Rank Revealing Decompositions ó Descomposiciones que Revelan el Rango). Se tratarán tanto los problemas de autovalores para matrices simétricas (SEVP, Symmetric Eigenvalue Problem) como los sistemas de ecuaciones lineales (LSE, Linear Systems of Equations) y los problemas de mínimos cuadrados (LSP, Least Square Problems). La importancia de la alta precisión relativa (HRA, High Relative Accuracy) viene principalmente de que se pueden construir algoritmos que permiten calcular las cantidades objetos de estudio (autovalores, solución de mínima longitud del LSP, ...) con la precisión que se merecen. En la última sección de este capítulo haremos un breve resumen de los algoritmos más relevantes de la HRA para el resto de esta memoria.

2.1. Notación

Denotaremos por $\|\cdot\|$ cualquier norma vectorial $\mathbb{R}^n$ y su correspondiente norma matricial inducida por $\|A\| := \max_{\|x\|=1} \|Ax\|$, a menos que en se indique en el subíndice. El símbolo I_n representará la matriz identidad de tamaño $n \times n$ y la

expresión A^T denotará la transpuesta de A. Dada una matriz simétrica $A \in \mathbb{R}^{n\times n}$, sus autovalores ordenados se denotarán por $\lambda(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ tales que $\lambda_i \geq \lambda_{i+1}$, para $i = 1 : n-1$. Dada $A \in \mathbb{R}^{m\times n}$, con $m \geq n$, sus valores singulares se denotarán por $\sigma_1(A) \geq \cdots \geq \sigma_n(A) \geq 0$. $\mathtt{u}$ es la unidad de redondeo del ordenador ($\mathtt{u} \approx 10^{-16}$). Dada una matriz $G \in \mathbb{R}^{m\times n}$ con entradas g_{ij}, $|G|$ denotará la matriz con entradas $|g_{ij}|$. Expresiones como $|G| \leq |B|$ significan que $|g_{ij}| \leq |b_{ij}|$ para $1 \leq i \leq m,\ 1 \leq j \leq n$. Denotaremos por $A^\dagger \in \mathbb{R}^{n\times m}$ la pseudoinversa de Moore-Penrose de $A \in \mathbb{R}^{m\times n}$. Recordemos que si $A \in \mathbb{R}^{n\times n}$ es no singular , entonces $A^\dagger = A^{-1}$, donde A^{-1} es la inversa de A. Usaremos la notación de MATLAB para submatrices: $A(i : j, :)$ indica la submatriz de A consistente de las filas i hasta la j y $A(:, k : l)$ indica la submatriz de A consistente de las columnas k hasta la l. El gap relativo en los autovalores lo notamos como $relgap_\lambda(A, \lambda_i)$, donde

$$relgap_\lambda(A, \lambda_i) := \min_{j\neq i} \frac{|\lambda_j - \lambda_i|}{|\lambda_i|}.$$

El gap relativo en los valores singulares lo representamos por $relgap_\sigma(A, \sigma_i)$, donde

$$relgap_\sigma(A, \sigma_i) := \min_{j\neq i} \frac{|\sigma_j - \sigma_i|}{\sigma_i}.$$

2.2. Teoría de perturbaciones del problema de autovalores y de valores singulares

Es bien conocido el teorema de perturbaciones de Weyl [69] que expresa el cambio en los autovalores bajo perturbaciones aditivas de una matriz simétrica[1].

Teorema 2.2.1 *Sean $\lambda_1 \geq \ldots \geq \lambda_n$ los autovalores de A y $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_n$ los autovalores de $A + E$, ambas simétricas, entonces, para $i = 1, \ldots, n$, se cumple que*

$$|\lambda_i - \widehat{\lambda}_i| \leq \|E\|_2 \tag{2.2.1}$$

$$\theta(v_i, \widehat{v}_i) \leq \frac{\|E\|_2}{\min_{j\neq i} |\lambda_i - \lambda_j|} \tag{2.2.2}$$

donde $\theta(v_i, \widehat{v}_i)$ es el ángulo agudo entre los autovectores v_i y $\widehat{v}_i$ correspondientes, respectivamente, a los autovalores λ_i y $\widehat{\lambda}_i$.

[1]Hemos añadido también por completitud y brevedad el resultado de autovectores, que no aparece habitualmente en el Teorema de Weyl.

Como se mencionó en la Introducción, estas cotas pueden no ser satisfactorias para los autovalores de módulos más pequeños, o para los correspondientes autovectores, o para autovectores con autovalores con gap pequeño $\text{mín}_{j\neq i}\,|\lambda_i - \lambda_j|$.

Sin embargo, si las perturbaciones son pequeñas pero en sentido multiplicativo, se puede demostrar a partir del Teorema de Weyl 2.2.1, que pequeñas perturbaciones multiplicativas de una matriz producen pequeñas perturbaciones relativas en sus autovalores y autovectores [37, 52].

Teorema 2.2.2 [37, Teorema 2.1] *Sea $A = A^T \in \mathbb{R}^{n\times n}$ y $\widetilde{A} = (I+E)A(I+E)^T \in \mathbb{R}^{n\times n}$, tal que $I+E$ es no singular. Además, supongamos que $\lambda_1 \geq \cdots \geq \lambda_n$ y $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_n$, son los autovalores de A y $\widetilde{A}$, respectivamente. Entonces*

$$|\widetilde{\lambda}_i - \lambda_i| \leq (2\,\|E\|_2 + \|E\|_2^2)\,|\lambda_i|, \quad \textit{para } i = 1, \ldots, n.$$

Teorema 2.2.3 [52, Teorema 3.1] *Sea $A = A^T \in \mathbb{R}^{n\times n}$ y $\widetilde{A} = (I+E)A(I+E)^T \in \mathbb{R}^{n\times n}$, tal que $I+E$ es no singular, con $\eta = \|E\| < 1$. Además, supongamos que $q_1 \geq \cdots \geq q_n$ y $\widetilde{q}_1 \geq \cdots \geq \widetilde{q}_n$, son los autovectores de A y $\widetilde{A}$, respectivamente. Entonces*

$$\sin\theta(q_i, \widetilde{q}_i) \leq \frac{\eta}{1-\eta}\left(1 + \frac{2+\eta}{relgap_{\widetilde{\lambda}}(A, \lambda_i)}\right), \qquad 1 \leq i \leq n,$$

donde $\theta(q_i, \widetilde{q}_i)$ es ángulo agudo entre q_i y $\widetilde{q}_i$.

Para el caso de autovalores múltiples, o clústers de autovalores muy cercanos en el sentido relativo, una cota similar se cumple para los senos de los ángulos canónicos de los correspondientes subespacios invariantes.

Un resultado similar se obtiene para la SVD:

Teorema 2.2.4 [37, Teorema 3.1], [52, Teorema 3.5] *Sean $A \in \mathbb{R}^{m\times n}$ y $\widetilde{A} = (I + E)\,A\,(I + F) \in \mathbb{R}^{m\times n}$ con SVDs, respectivamente, $A = U\Sigma V^T$, $\widetilde{A} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$. Sean $\eta_E = \|E\|_2, \eta_F = \|F\|_2, \eta = \text{máx}\{\eta_E, \eta_F\}$ y $\eta' = 2\eta + \eta^2$.*

Entonces

1. *La diferencia entre los valores singulares de A y $\widetilde{A}$ obedece*

$$\frac{|\sigma_i - \widetilde{\sigma}_i|}{\sigma_i} \leq \eta_E + \eta_F + \eta_E\eta_F \leq \eta', \qquad 1 \leq i \leq n.$$

2. *El ángulo entre los vectores singulares izquierdos u_i y $\widetilde{u}_i$ (o entre los derechos v_i y $\widetilde{v}_i$) obedece*

$$\sin\theta_i \leq \sqrt{2}\left(\frac{1+\eta'}{1-\eta'}\frac{\eta'}{\min(relgap_{\widetilde{\sigma}}(A,\sigma_i),2)-\eta'}+\eta\right), \qquad 1\leq i\leq n,$$

siempre que $\min(relgap_{\widetilde{\sigma}}(A,\sigma_i),2) > \eta'$. *En el caso de valores singulares múltiples, o grupos de valores singulares muy próximos, hay una cota similar para los senos de los ángulos canónicos entre los subespacios correspondientes.*

Los Teoremas 2.2.2, 2.2.3 y 2.2.4 son la base para los análisis de HRA para los problemas SEVP y SVD. La técnica siempre pasa por expresar el análisis de errores como un resultado multiplicativo regresivamente y después aplicar los teoremas anteriores, esto se verá en la Sección 2.5 y en el Capítulo 4.

2.3. Teoría de perturbaciones de sistemas de ecuaciones lineales y problemas de mínimos cuadrados

En esta sección presentamos resultados conocidos de teoría de perturbaciones para sistemas de ecuaciones lineales y problemas de mínimos cuadrados, teniendo en cuenta que los resultados de perturbaciones aditivas producen cambios relativos muy grandes en la solución del problema, por lo tanto, si nuestro interés es obtener alta precisión relativa, tendremos que buscar una alternativa para perturbar adecuadamente el problema. Para los sistemas de ecuaciones lineales y el problema de mínimos cuadrados, procederemos de la siguiente manera, analizamos el efecto que tiene en la solución el hecho de perturbar sólamente el lado derecho del problema y observamos que el número de condición asociado a esta particular estructura de perturbaciones es pequeño. Luego perturbamos adecuadamente ambos lados del problema, es decir perturbamos multiplicativamente la matriz de coeficientes asociada y perturbamos el vector del lado derecho. Observamos que las cotas relativas conseguidas son pequeñas.

Teorema 2.3.1 [43, Teorema 7.2] *Sea* $Ax = b$ *y* $(A+\Delta A)\widetilde{x} = b+h$, *donde* $\|\Delta A\| \leq \epsilon\|E\|$ *y* $\|h\| \leq \epsilon\|b\|$. *Además, supongamos que* $\epsilon\|A^{-1}\|\|E\| < 1$. *Entonces*

$$\frac{\|\widetilde{x}-x\|}{\|x\|} \leq \frac{\epsilon}{1-\epsilon\|A^{-1}\|\|E\|}\left(\frac{\|A^{-1}\|\|b\|}{\|x\|}+\|A^{-1}\|\|E\|\right).$$

El número de condición ($\frac{\|A^{-1}\|\|f\|}{\|x\|}+\|A^{-1}\|\|E\|$) y la sensibilidad del problema aparecen generalmente como lo indicamos en el Teorema 2.3.1, pero en casos especiales, el número de condición no aparece completo. Por ejemplo, si queremos calcular la solución del sistema de ecuaciones lineales $Ax = b$, y consideramos perturbaciones sólamente en el vector b, es decir, consideramos que $A\widetilde{x} = b + h$ con $\|h\| \leq \epsilon\|b\|$, entonces: $\widetilde{x} = A^{-1}(b+h)$, con lo que tenemos un resultado para la sensibilidad de la solución de un sistema de ecuaciones lineales respecto a perturbaciones en el vector b.

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|h\|}{\|x\|} \leq \epsilon\frac{\|A^{-1}\|\|b\|}{\|x\|}, \tag{2.3.1}$$

donde $\frac{\|A^{-1}\|\|b\|}{\|x\|}$ es el número de condición de $A\widetilde{x} = (b + h)$ con $\|h\| \leq \epsilon\|b\|$. Particularmente, se ha demostrado en [43, pág. 121] que

$$\kappa(A, b) = \lim_{\epsilon\to 0} \sup\left\{ \frac{\|\widetilde{x} - x\|}{\epsilon\|x\|} : A\,\widetilde{x} = b + h, \ \ \|h\| \leq \epsilon\,\|b\| \right\} = \frac{\|A^{-1}\|\,\|b\|}{\|x\|}. \tag{2.3.2}$$

Como $\kappa(A) = \|A^{-1}\|\|A\|$ y $Ax = b$, entonces,

$$1 \leq \kappa(A, b) \leq \kappa(A),$$

pero la idea es demostrar que si fijamos A de tal manera que $\kappa(A) \gg 1$, entonces $\kappa(A, b) \ll \kappa(A)$ para la mayoría de vectores b, es decir que, $\kappa(A, b)$ usualmente es pequeño. La explicación de este hecho para el caso de la $\|\cdot\|_2$ y su norma matricial inducida, lo presentamos a continuación.

Teorema 2.3.2 [14] *Sea $A = U\Sigma V^T$ la SVD de $A \in \mathbb{R}^{m\times n}$ tal que $rango(A) = r \leq n \leq m$ y P_k la proyección ortogonal sobre el subespacio generado por las últimas k columnas de U. Entonces*

$$\kappa_2(A, b) \leq \frac{\sigma_{r+1-k}}{\sigma_r}\frac{\|b\|_2}{\|P_k b\|_2}, \quad \textit{para} \quad k = 1 : r. \tag{2.3.3}$$

Si $\cos\theta(u_r, b) \approx 0$, entonces $\|P_2 b\|_2 \approx |u_{r-1}^T b|$, donde u_{r-1} es la penúltima columna de U,

$$\kappa_2(A, b) \lesssim \frac{\sigma_{r-1}}{\sigma_r}\frac{1}{\cos\theta(u_{r-1}, b)}.$$

Si $\sigma_{r-1} \approx \sigma_r$, esta cota será pequeña con un alta probabilidad para vectores aleatorios b tales que $\cos\theta(u_r, b) \approx 0$. Observemos que (2.3.3) puede interpretarse diciendo que el número de condición efectivo para el problema $A\widetilde{x} = b+h$ es σ_{r+1-k}/σ_r, donde k es un número natural pequeño, tal que $\|P_k b\|_2/\|b\|_2$ no es muy pequeño.

La teoría de perturbaciones multiplicativas ha recibido considerable atención cuando se desea obtener soluciones precisas del problema de autovalores y del problema de valores singulares [37, 45, 46, 51, 52]. Para el problema de encontrar soluciones precisas de sistemas de ecuaciones la teoría de perturbaciones multiplicativas apareció recientemente en [32, Lema 3.1], sin embargo presentamos una versión a primer orden en ϵ de este resultado.

Teorema 2.3.3 [32, Lema 3.1] *Sea* $Ax = b$ *y* $(I + E)A(I + F)\widetilde{x} = b + h$, *donde* $\|h\| \leq \epsilon\|b\|$, $I + E \in \mathbb{R}^{n\times n}$ *y* $I + F \in \mathbb{R}^{n\times n}$ *son matrices no singulares tales que* $\max\{\|E\|, \|F\|\} \leq \epsilon < 1$. *Entonces, a primer orden en* ϵ *se cumple que*

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \epsilon \left(1 + 2\frac{\|A^{-1}\|\|b\|}{\|x\|}\right) + O(\epsilon^2). \tag{2.3.4}$$

Observemos que el factor $\frac{\|A^{-1}\|\|b\|}{\|x\|}$ que aparece en la cota (2.3.4) es pequeño, con lo que, pequeñas perturbaciones multiplicativas en la matriz implican pequeñas perturbaciones en la solución. Como una conclusión importante, tenemos que el factor $\frac{\|A^{-1}\|\|b\|}{\|x\|}$ no aparece sólamente en perturbaciones del lado derecho, sino que aparece en perturbaciones de tipo multiplicativo.

Siguiendo las mismas líneas presentadas en esta sección para sistemas de ecuaciones, presentamos resultados análogos para el problema de mínimos cuadrados $Ax \approx b$.

Teorema 2.3.4 [43, Teorema 20.1] *Sean* $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ *y* $(A + \Delta A)$ *con* $\|\Delta A\|_2 \leq \epsilon\|A\|_2$ *matrices de rango completo. Supongamos además que*

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2, \quad y \quad r = b - Ax. \tag{2.3.5}$$

Entonces, si $\epsilon\,\kappa_2(A) < 1$ *tenemos que*

$$\frac{\|\widetilde{x} - x\|_2}{\|x\|_2} \leq \frac{\epsilon\,\kappa_2(A)}{1 - \epsilon\,\kappa_2(A)} \left(2 + (\kappa_2(A) + 1)\frac{\|r\|_2}{\|A\|_2\|x\|_2}\right) \tag{2.3.6}$$

Si escribimos el Teorema 2.3.4 a primer orden en ϵ, obtenemos el siguiente resultado.

Teorema 2.3.5 *Sean* $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ *y* $(A+\Delta A)$ *con* $\|\Delta A\|_2 \leq \epsilon\|A\|_2$ *matrices de rango completo. Supongamos además que*

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2, \quad y \quad r = b - Ax. \tag{2.3.7}$$

Entonces, si $\epsilon\,\kappa_2(A) < 1$ *tenemos que*

$$\frac{\|\widetilde{x}-x\|_2}{\|x\|_2} \leq \epsilon\,\kappa_2(A) + \epsilon\,\kappa_2^2(A)\frac{\|r\|_2}{\|A\|_2\|x\|_2} + \epsilon\,\kappa_2(A)\frac{\|b\|_2}{\|A\|_2\|x\|_2} + O(\epsilon^2). \tag{2.3.8}$$

Si resolvemos el problema de mínimos cuadrados $A\widetilde{x} \approx b + h$ con $\|h\|_2 \leq \epsilon\|b\|_2$, entonces la solución $\widetilde{x}$ es tal que: $\widetilde{x} = A^\dagger(b+h) = x + A^\dagger h$, y por lo tanto

$$\frac{\|\widetilde{x}-x\|_2}{\|x\|_2} \leq \frac{\|A^\dagger\|_2\|h\|_2}{\|x\|_2} \leq \epsilon\frac{\|A^\dagger\|_2\|b\|_2}{\|x\|_2}, \tag{2.3.9}$$

donde $\frac{\|A^\dagger\|\|b\|}{\|x\|}$ es el número de condición del problema de mínimos cuadrados perturbando sólo el vector b. En la Sección 3.3.1, demostramos que este factor habitualmente es pequeño.

Por el análisis realizado para sistemas de ecuaciones, sabemos que si perturbamos multiplicativamente la matriz podemos obtener que el número de condición del problema es $\frac{\|A^\dagger\|\|b\|}{\|x\|}$, pero como este factor es habitualmente pequeño, entonces tenemos que pequeñas perturbaciones multiplicativas en los factores implican pequeñas perturbaciones multiplicativas en la solución. Esto lo resumimos en el siguiente teorema, el cual es un resultado a primer orden del Teorema 3.3.1.

Teorema 2.3.6 *Sea* $Ax \approx b$ *y* $(I+E)A(I+F)\widetilde{x} \approx b + h$, *donde* $\|h\|_2 \leq \epsilon\|b\|_2$, *y* $\text{máx}\{\|E\|_2, \|F\|_2\} \leq \epsilon < 1$. *Entonces, a primer orden en* ϵ *se cumple que*

$$\frac{\|\widetilde{x}-x\|_2}{\|x\|_2} \leq \epsilon\left(1 + 4\frac{\|A^\dagger\|_2\|b\|_2}{\|x\|_2}\right) + O(\epsilon^2). \tag{2.3.10}$$

2.4. Descomposiciones que revelan el rango y teoría de perturbaciones

Como ya se ha indicado en la Introducción, en [20] se presentó un tratamiento unificado para el cálculo de la SVD con alta precisión relativa mediante el uso de la *RRD* (*Rank Revealing Decomposition*). Partiendo de ese resultado para valores singulares, con posterioridad se ha ido aplicando esa técnica a los otros tres problemas del Álgebra Lineal: problema de autovalores [30] y Capítulo 4, sistemas de ecuaciones lineales [32] y por último problema de mínimos cuadrados que corresponde al Capítulo 3. En esta línea, una de las ideas fundamentales en este trabajo son los conceptos de *RRD* y *RRD* precisa [20] (ver también [43, Sección 9.12]).

Definición 2.4.1 *Sea $A \in \mathbb{C}^{m\times n}$, la factorización XDY de A, donde $X \in \mathbb{C}^{m\times r}$, $D = \mathrm{diag}(d_1, d_2, \ldots, d_r) \in \mathbb{C}^{r\times r}$ es diagonal y no singular e $Y \in \mathbb{C}^{r\times n}$, son tales que $rango\,(X) = rango\,(Y) = r$, y X e Y son matrices bien condicionadas, la llamaremos RRD de A.*

Notemos que esto significa que el rango de A es r, y que si A está mal condicionada, entonces el factor diagonal D es también mal condicionado.

Para obtener soluciones precisas en el marco teórico que introducimos en este trabajo es necesario partir de RRDs precisas, por lo tanto, definimos la RRD precisa de una matriz A siguiendo las líneas de la definición dada en [20].

Definición 2.4.2 *Sea $A = XDY$ una RRD de $A \in \mathbb{C}^{m\times n}$, donde $X \in \mathbb{C}^{m\times r}$, $D = \mathrm{diag}(d_1, \ldots, d_r) \in \mathbb{C}^{r\times r}$ e $Y \in \mathbb{C}^{r\times n}$ y sean $\widehat{X} \in \mathbb{C}^{m\times r}$, $\widehat{D} = \mathrm{diag}(\widehat{d}_1, \ldots, \widehat{d}_r) \in \mathbb{C}^{r\times r}$ e $\widehat{Y} \in \mathbb{C}^{r\times n}$ los factores calculados por un cierto algoritmo en un ordenador con unidad de redondeo $\mathtt{u}$. Decimos que la factorización $\widehat{X}\widehat{D}\widehat{Y}$ ha sido calculada de manera precisa, o es una RRD precisa, si*

$$\frac{\|\widehat{X} - X\|_2}{\|X\|_2} \leq \mathtt{u}\, p(m,n), \quad \frac{\|\widehat{Y} - Y\|_2}{\|Y\|_2} \leq \mathtt{u}\, p(m,n) \quad y \tag{2.4.1}$$

$$\frac{|\widehat{d}_i - d_i|}{|d_i|} \leq \mathtt{u}\, p(m,n), \quad i = 1:r, \tag{2.4.2}$$

donde $p(m,n)$ es una función moderadamente creciente de m y n, es decir, una función acotada por un polinomio de grado bajo en m y n, tal que

$$\mathrm{máx}\{\kappa_2(X)\,,\, \kappa_2(Y)\}\, \mathtt{u}\, p(m,n) < 1/2\,. \tag{2.4.3}$$

2.4.1. Descomposición en valores singulares y problema simétrico de autovalores

El resultado fundamental de [20] se basaba en que los factores de una RRD determinan de manera precisa la SVD de la matriz original, es decir, pequeñas perturbaciones relativas componente a componente en los elementos de la matriz D y pequeñas perturbaciones relativas en norma en X e Y, producen pequeñas perturbaciones en los valores singulares de A y pequeñas perturbaciones en los vectores singulares con respecto al gap relativo de los valores singulares.

Teorema 2.4.3 [20, Teorema 2.1] *Sean* $A = XDY^T$ *y* $\widetilde{A} = \widetilde{X}\widetilde{D}\widetilde{Y}^T$ *RRDs con SVDs, respectivamente,* $A = U\Sigma V^T$, $\widetilde{A} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$, *donde* $\widetilde{X}, \widetilde{D}$ *e* $\widetilde{Y}^T$ *están definidas así:*

$$\begin{aligned}
\widetilde{X} &= X + \delta X, \quad con \quad \frac{\|\delta X\|_2}{\|X\|_2} \leq \epsilon \\
\widetilde{Y} &= Y + \delta Y, \quad con \quad \frac{\|\delta Y\|_2}{\|Y\|_2} \leq \epsilon \\
\widetilde{D} &= D + \delta D, \quad con \quad \frac{|\delta D_{ii}|}{|D_{ii}|} \leq \epsilon \quad para\ todo\ \ i,
\end{aligned}$$

con $0 \leq \epsilon < 1$. *Sea* $\eta = \epsilon\,(2+\epsilon)\ \max\{\kappa_2(X), \kappa_2(Y)\} < 1$ *y* $\eta' = 2\eta + \eta^2$. *Entonces*

1. $\widetilde{A} = (I+E)\ A\ (I+F)$ *con* $\max\{\|E\|_2, \|F\|_2\} \leq \eta$.

2. *La diferencia entre los valores singulares de* A *y* $\widetilde{A}$ *obedece*

$$|\sigma_i - \widetilde{\sigma}_i| \ \leq\ \eta'\,|\sigma_i|, \qquad 1 \leq i \leq n.$$

3. *El ángulo entre los vectores singulares izquierdos* u_i *y* $\widetilde{u}_i$ *(o entre los derechos* v_i *y* $\widetilde{v}_i$*) obedece*

$$\sin\theta_i \leq \sqrt{2}\left(\frac{1+\eta'}{1-\eta'}\frac{\eta'}{\min(relgap_\sigma(A,\sigma_i),2)-\eta'} + \eta\right), \qquad 1 \leq i \leq n,$$

siempre que $\min(relgap_\sigma(A,\sigma_i),2) > \eta'$. *En el caso de valores singulares múltiples, o grupos de valores singulares muy próximos, hay una cota similar para los senos de los ángulos canónicos entre los subespacios correspondientes.*

Hemos modificado ligeramente el enunciado que se presenta en [20, Teorema 2.1] añadiendo el punto 1 de los resultados. Queremos destacar de forma explícita que bajo las hipótesis del teorema, la perturbación de los factores se puede escribir como una pequeña perturbación multiplicativa de la matriz. Los otros dos resultados del teorema se obtienen inmediatamente sin más que aplicar el teorema 2.2.4. De esta manera destacamos el papel crucial de la teoría multiplicativa de perturbaciones en establecer, en este caso, que valores y vectores singulares son poco sensibles a perturbaciones de la RRD.

De una forma similar al problema de valores singulares anterior se puede tratar el problema simétrico de autovalores. Dopico y Koev demostraron en [28] también que a partir de una RRD simétrica y precisa de una matriz también simétrica es posible calcular sus autovalores y autovectores con alta precisión relativa.

Teorema 2.4.4 [28, Teorema 2.1] *Sean* $A = XDX^T$ *y* $\widetilde{A} = \widetilde{X}\widetilde{D}\widetilde{X}^T$ *RRDs de las matrices simétricas* $A \in \mathbb{R}^{n\times n}$ *y* $\widetilde{A} \in \mathbb{R}^{n\times n}$ *respectivamente. Sean* $\lambda_1 \geq \cdots \geq \lambda_n$ *los autovalores de* A *y* $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_n$ *los autovalores de* $\widetilde{A}$*. Sean* $q_1, \ldots, q_n$ *y* $\widetilde{q}_1, \ldots, \widetilde{q}_n$ *los correspondientes autovectores ortonormales. Supongamos que*

$$\frac{\|\widetilde{X} - X\|_2}{\|X\|_2} \leq \beta,$$
$$\frac{|\widetilde{D}_{ii} - D_{ii}|}{|D_{ii}|} \leq \beta \quad \textit{para todo } i,$$

donde $0 \leq \beta < 1$*. Sea* $\eta = \beta(2+\beta)\kappa_2(X)$ *menor que 1; entonces*

$$|\lambda_i - \widetilde{\lambda}_i| \leq (2\eta + \eta^2)|\lambda_i|, \qquad 1 \leq i \leq n,$$

y

$$\sin\theta(q_i, \widetilde{q}_i) \leq \frac{\eta}{1-\eta}\left(1 + \frac{2+\eta}{relgap_{\widetilde{\lambda}}(A, \lambda_i)}\right), \qquad 1 \leq i \leq n,$$

donde $\theta(q_i, \widetilde{q}_i)$ *es el ángulo agudo entre* q_i *y* $\widetilde{q}_i$*. Para el caso de autovalores múltiples, o clústers de autovalores muy cercanos en el sentido relativo, una cota similar se cumple para los senos de los ángulos canónicos de los correspondientes subespacios invariantes.*

2.4.2. Sistemas de ecuaciones lineales vía la RRD

Las técnicas anteriores también se han extendido a los otros dos grandes problemas del *Álgebra Lineal Numérica*, los sistemas de ecuaciones lineales [32] y los problemas de mínimos cuadrados [11]. Este último problema es el objeto del Capítulo 3 de esta memoria, y allí nos ocuparemos de él.

En [32] se demostró que si la matriz de un LSE $Ax = b$ se puede factorizar de forma precisa entonces la solución del LSE cambia poco al variar los factores de la RRD. Se presenta aquí sólo el resultado a primer orden (para el resultado completo ver [32, Teorema 3.2])

Teorema 2.4.5 [32, Teorema 3.2] *Sea* $A \in \mathbb{C}^{n\times n}$ *con RRD* XDY *y* $b \in \mathbb{C}^n$*. Sea* $XDYx = b$ *y* $(X + \Delta X)(D + \Delta D)(Y + \Delta Y)\widetilde{x} = b + h$*, donde* $\|\Delta X\| \leq \epsilon\|X\|$*,* $\|\Delta Y\| \leq \epsilon\|Y\|$ $|\Delta D| \leq \epsilon|D|$ *y* $\|h\| \leq \epsilon\|b\|$ *y supongamos que* $\epsilon\kappa(Y) < 1$ *y que* $\epsilon(2+\epsilon)\kappa(X) < 1$*. Entonces, a primer orden en* ϵ*,*

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \epsilon\left(\kappa(Y) + \left(1 + 2\,\|X\|\frac{\|X^{-1}b\|}{\|b\|}\right)\,\frac{\|A^{-1}\|\,\|b\|}{\|x\|}\right) + O(\epsilon^2). \qquad (2.4.4)$$

O, en una forma más débil, pero de mayor utilidad en la práctica:

$$\frac{\|\widetilde{x}-x\|}{\|x\|} \leq \epsilon \left(\kappa(Y) + (1+2\,\kappa(X))\ \frac{\|A^{-1}\|\,\|b\|}{\|x\|}\right) + O(\epsilon^2),$$

ya que no requieren que se calcule $\|X^{-1}b\|$ *y no sobreestima significativamente la variación de la solución si* $\kappa(X)$ *es pequeño.*

Ya se presentó en la Sección 2.3 una breve explicación de por qué el factor $\frac{\|A^{-1}\|\,\|b\|}{\|x\|}$ es habitualmente pequeño. En la Sección 3.3.1 se presentará el factor equivalente para el LSP $\frac{\|A^{\dagger}\|\,\|b\|}{\|x\|}$ y se verá también que es usualmente pequeño.

2.5. Algoritmos y errores

El estudio de la alta precisión relativa tiene su fin último en poder *calcular* magnitudes con más precisión que la que garantizan los algoritmos convencionales. En esta sección vamos a describir los algoritmos presentes en la literatura que garantizan HRA. Describiremos los algoritmos y los resultados que se obtienen de sus análisis de errores, es decir la precisión esperada cuando se ejecutan en un ordenador.

Antes de empezar recordaremos que en los análisis de errores presentados en la Secciones 3.5 y 4.3 usaremos el modelo de errores convencional para aritmética en coma flotante [43, Sección 2.2]

$$fl(a \odot b) = (a \odot b)(1+\delta),$$

donde a y b son números reales en coma flotante, $\odot \in \{+,-,\times,/\}$, y $|\delta| \leq \mathtt{u}$. Recordemos que este modelo también se cumple para números complejos en coma flotante si $\mathtt{u}$ es reemplazada por una constante ligeramente más grande, ver [43, Sección 3.6]. Suponemos también que no ocurre *overflow* ni *underflow*.

El campo de la HRA tiene una larga tradición (veánse las referencias citadas en el Capítulo 1), sin embargo en la forma en que se enfoca en esta memoria data de los primeros años de la década de 1990. Entre las referencias destacamos la de Demmel y Veselić [25]. En ella se inicia el esquema, que ha resultado tan fructífero después, de Factorización+Algoritmo específico para lograr la HRA. En ese trabajo se mostraba que el algoritmo de Jacobi implícito aplicado a los factores de Cholesky de una matriz simétrica definida positiva podía alcanzar HRA, al contrario que el algoritmo QR; esto también sirvió para traer a primera plana de los algoritmos del Álgebra Lineal Numérica los algoritmo de tipo Jacobi que estaban ligeramente olvidados.

Los algoritmos de tipo Jacobi están basados en las rotaciones del mismo nombre.

$$R(i,j,c,s) = \begin{array}{c} \\ i \\ \\ j \\ \\ \end{array} \begin{array}{c} \begin{array}{cccccc} & & i & & j & \end{array} \\ \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & c & & -s & \\ & & & \ddots & & \\ & & s & & c & \\ & & & & & \ddots \\ & & & & & & 1 \end{bmatrix} \end{array}, \tag{2.5.1}$$

donde los cosenos c, y los senos s, se eligen en cada caso para que la caja 2×2 adecuada se diagonalice:

$$\begin{bmatrix} c & -s \\ s & c \end{bmatrix}^T \begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix} \begin{bmatrix} c & -s \\ s & c \end{bmatrix} = \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{bmatrix}.$$

Para más detalles ver los siguientes libros clásicos [17, Sección 5.3.5], [39, Sección 8.5] y [61, Capítulo 9].

Quizás la forma más conocida de usar las rotaciones de Jacobi es para diagonalizar una matriz simétrica haciendo rotaciones simultáneas, por la derecha y por la izquierda. Es lo que se conoce como el algoritmo de Jacobi "two-sided". Ligeramente menos conocido es el algoritmo de Jacobi "one-sided"que sirve para calcular la SVD de una matriz (ver por ejemplo [39]).

La esencia del algoritmo "one-sided" es aplicar implícitamente el algoritmo "two-sided" a la matriz $G = A^T A$. Ésta no se calcula nunca explícitamente, según se van necesitando sus elementos se calculan con la expresión:

$$g_{ij} = \sum_{k=1}^{n} a_{ki} a_{kj} \tag{2.5.2}$$

Algoritmo 2.5.1 (Algoritmo de Jacobi "one-sided")

`Input:` $A \in \mathbb{R}^{n \times n}$

`Output:` $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$ y $V \in \mathbb{R}^{n \times n}$ matrices de valores y vectores, singulares derechos respectivamente.

$V = I$

repetir

para i, j

calcular g_{ii}, g_{ij}, g_{jj} de $G = A^T A$ como en la ecuación (2.5.2)

calcular una rotación de Jacobi $R(i, j, c, s)$ tal que:

$$R^T(i, j, c, s) \begin{bmatrix} g_{ii} & g_{ij} \\ g_{ij} & g_{jj} \end{bmatrix} R(i, j, c, s) = \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{bmatrix}$$

$A = A\, R(i, j, c, s)$

$V = V\, R(i, j, c, s)$

fin para

hasta convergencia $\frac{|g_{ij}|}{\sqrt{|g_{ii} g_{jj}|}} < tol$

calcular $\sigma_i = \|G(:, i)\|$ para $i = 1, \ldots, n.$

los vectores singulares izquierdos, serán las columnas

normalizadas de G.

A partir del algoritmo "one-sided" 2.5.1 para calcular la SVD, se puede diseñar el siguiente algoritmo [25] para calcular los autovalores y los autovectores de matrices simétricas definidas positivas:

Algoritmo 2.5.2 *[25, Algoritmo 4.4]*

Input: $A = A^T \in \mathbb{R}^{n \times n}$ definida positiva.

Output: Autovalores λ_i y autovectores Q de A.

Paso 1: Calcular una descomposición de Cholesky con pivote completo de $A = L\, L^T$.

Paso 2: Calcular la SVD de L usando el Algoritmo 2.5.1: $LR_1R_2 \cdots R_p = U\Sigma$, donde $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$.

Paso 3: $\lambda_i = \sigma_i^2$ para $i = 1 : n$ y $Q = U$.

Obsérvese que es un algoritmo de tipo *2 pasos* como los usados en esta memoria y en otras situaciones (ver el Capítulo 1).

Se demostró en [25, 55, 56] que si los autovalores y los autovectores de matrices simétricas definidas positivas son calculados por el algoritmo 2.5.2, entonces, los errores en los autovalores calculados están dados por el siguiente resultado.

Teorema 2.5.3 *Sea $A = SBS$ una matriz simétrica definida positiva, con $S = diag[a_{11}^{1/2}, \ldots, a_{nn}^{1/2}]$, $b_{ii} = 1$ y $\{\lambda_i\}_{i=1}^n$ los autovalores de A. Sean $\{\widehat{\lambda}_i\}_{i=1}^n$ los autovalores calculados por el Algoritmo 2.5.2 en aritmética finita con precisión* $\mathtt{u}$*. Entonces*

$$\frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq O(\mathtt{u}) \left[\frac{1}{\lambda_n(B)} + \frac{1}{\sqrt{\lambda_n(B)}} \right] \tag{2.5.3}$$

Si la matriz A está bien escalada, es decir, si $\kappa_2(B) \gtrsim 1$, entonces todos los autovalores están calculados con varios dígitos de precisión.

El siguiente hito en la historia reciente de los algoritmos de alta precisión relativa, es el Algoritmo 3.1 de [20], que calcula la solución precisa del problema de valores singulares de A, vía la RRD de A.

Algoritmo 2.5.4 (Solución precisa del problema de valores singulares)

`Input:` $X \in \mathbb{R}^{m\times n}, D \in \mathbb{R}^{n\times n}$ e $Y \in \mathbb{R}^{n\times n}$ tales que $A = XDY$ es una RRD de A.

`Output:` $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$ la matriz de valores singulares y, $U \in \mathbb{R}^{m\times n}$ y $V \in \mathbb{R}^{n\times n}$ matrices de vectores singulares izquierdos y derechos respectivamente.

`Paso 1:` Calcular la factorización QR con pivote por columnas $XD = QRP$.

`Paso 2:` Multiplicar las matrices R, P e Y para obtener $W = RPY$

`Paso 3:` Calcular la SVD de $W = \overline{U}\Sigma V^T$ usando el Algoritmo 2.5.1.

`Paso 4:` multiplicar las matrices Q y $\overline{U}$ para obtener $U = Q\overline{U}$

La alta precisión relativa la garantizamos a partir de la teoría multiplicativa de perturbaciones, Teorema 2.4.3, y el análisis de errores del Algoritmo:

Teorema 2.5.5 [33, Teorema 2.1] *El Algoritmo 2.5.4 produce errores multiplicativos regresivos cuando es ejecutado con unidad de redondeo* $\mathtt{u}$*, es decir, Si $A = XDY \in \mathbb{R}^{m\times n}$ es la RRD calculada en el paso 1 del Algoritmo 2.5.4 y $\widehat{U}\widehat{\Sigma}\widehat{V}^T$ es la SVD calculada por el algoritmo, entonces, existen matrices $U' \in \mathbb{R}^{m\times r}, V' \in \mathbb{R}^{n\times r}, E \in \mathbb{R}^{m\times m}$ y $F \in \mathbb{R}^{n\times n}$ tales que U' y V' tienen columnas ortonormales,*

$$\|U' - \widehat{U}\| = O(\mathtt{u}), \quad \|V' - \widehat{V}\| = O(\mathtt{u})$$
$$\|E\| = O(\mathtt{u}\kappa(X)), \quad \|F\| = O(\mathtt{u}\kappa(R')\kappa(Y)),$$

donde R' es la matriz con mejor número de condición de todos los escalamientos por filas de la matriz triangular R que aparece en el Algorithm 3.1 de [20] y

$$(I+E)A(I+F) = U'\widehat{\Sigma}V'^T.$$

Se demostró en [20], que $\kappa(R')$ es a lo sumo de orden $O(n^{3/2}\kappa(X))$, pero en la práctica se observa mediante diferentes experimentos numéricos que $\kappa(R')$ se comporta como $O(n)$.

A continuación presentamos otro hito significativo. El algoritmo de Jacobi implícito [30, Algoritmo 1] aplicado a matrices simétricas, definidas o indefinidas, produce HRA si se parte de una matriz de la que se ha calculado una RRD precisa.

Sea $A = X\Omega X^T$ una factorización simétrica de A, donde $X \in \mathbb{R}^{n\times n}$ y $\Omega = diag(\omega_1, \ldots, \omega_n) \in \mathbb{R}^{n\times n}$ son matrices no singulares. La idea es aplicar el algoritmo simétrico ("two-sided") de Jacobi de forma implícita, eligiendo en cada paso una rotación de Jacobi R tal que la posición (i,j) de $A = X\Omega X^T$ sea cero. En cada paso del proceso sólo se actualiza el factor X y la matriz Ω, que es la que porta el mal condicionamiento, permanece constante, es decir, pasamos de $X\Omega X^T$ a $R^T(X\Omega X^T)R$, manteniendo la matriz factorizada ($X \to R^T X$). Se demostró en [30, Teorema 6] que si $X\Omega X^T$ es una descomposición precisa que revele el rango, entonces el Algoritmo de Jacobi implícito, calcula soluciones precisas del problema de autovalores y autovectores.

Algoritmo 2.5.6 (Algoritmo de Jacobi implícito [30, Algoritmo 1])

`Input:` $X \in \mathbb{R}^{n\times n}$ una matriz bien condicionada y $\Omega = diag(\omega_1, \ldots, \omega_n) \in \mathbb{R}^{n\times n}$ no singular.

`Output:` $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ y $U \in \mathbb{R}^{n\times n}$ matrices de autovalores y autovectores, respectivamente.

$\widehat{\kappa}(X)$ `es el valor calculado de` $\kappa(X)$
$U = I_n$
$G = X\ diag(\sqrt{|\omega_1|}, \ldots, \sqrt{|\omega_n|})$
$J = diag(sign(\omega_1), \ldots, sign(\omega_n))$
`repetir`
`para` $i = 1$ `hasta` $n-1$
    `para` $j = i+1$ `hasta` n
        `calcular` a_{ii}, a_{ij}, a_{jj} `de` $A = GJG^T$ `usando`
            $a_{ij} = \sum_{k=1}^{n} g_{ik}g_{jk}\, sign(\omega_k)$

```
            calcular
            T = [c  -s; s  c],  c^2 + s^2 = 1 tal que  T^T [a_ii a_ij; a_ij a_jj] T = [μ1 0; 0 μ2]
            G = R(i, j, c, s)^T G
            U = U R(i, j, c, s)^2
        fin para
    fin para
    hasta convergencia ( |a_ij| / sqrt(|a_ii a_jj|) ≤ ε máx{n, κ̂(X)}  para todo i < j  y
                         (Σ_{k=1}^n g_ik^2) / |a_ii| ≤ 2κ̂(X)  para todo i )
    calcular  λ_i = a_ii  para  i = 1, ..., n.
```

Las cotas de errores multiplicativas regresivas dadas en el Teorema 2.5.7 pueden combinarse fácilmente con los Teoremas 2.2.2 y 2.2.3 para probar que el Algoritmo 2.5.6 calcula los autovalores y los autovectores de una RRD XDX^T con alta precisión relativa:

Teorema 2.5.7 [30, Teorema 6] *Si N_R es el número de rotaciones de Jacobi aplicadas en el Algoritmo 2.5.6 hasta que se satisfaga el criterio de parada, $\widehat{\kappa}(X)\gamma_{n+1} < \frac{1}{8}$, y $\sqrt{n}\kappa(X)\gamma_2 < \frac{1}{2}$, entonces la matriz de autovalores calculada, $\widehat{\Lambda} = \mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_n)$, y la matriz calculada de autovectores, $\widehat{U}$, están cerca de las matrices de autovalores y autovectores de una pequeña perturbación multiplicativa de XDX^T. Es decir, existe una matriz ortogonal $U \in \mathbb{R}^{n\times n}$ tal que*

$$U\widehat{\Lambda}U^T = (I+E)\,XDX^T\,(I+E)^T, \tag{2.5.4}$$

con $\|E\|_F = O(\mathtt{u}\,(n^2\widehat{\kappa}(X) + N_R\kappa(X)))$ y $\|\widehat{U} - U\|_F = O(N_R\,\mathtt{u})$.

Por último, extendiendo las ideas de usar la RRD como buena parametrización que refleja la sensibilidad de los problemas espectrales a los sistemas de ecuaciones lineales, Dopico y Molera [32] mostraron cómo se pueden calcular soluciones con alta precisión de LSE, $Ax = b$, para cualquier matriz de la que se pueda obtener una RRD precisa mediante el siguiente algoritmo:

Algoritmo 2.5.8 *[32, Algoritmo 4.1]*
`Input:` $A \in \mathbb{C}^{n\times n}$, $b \in \mathbb{C}^n$

[2]Ver (2.5.1), $R(i,j,c,s)$ es la notación usual para las rotaciones de Jacobi en los índices (i,j) y en el ángulo θ, es decir: $R(i,j,c,s) = I_n$ excepto en las posiciones $R(i,j,c,s)_{ii} = R(i,j,c,s)_{jj} = \cos(\theta) = c$ y $R(i,j,c,s)_{ij} = -R(i,j,c,s)_{ji} = -\sin(\theta) = -s$.

`Output:` x, *solución de* $A\,x = b$

`Paso 1:` *Calcular una RRD precisa de* $A = X\,D\,Y$, *con* $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$.

`Paso 2:` *Resolver los tres sistemas de ecuaciones,*

$$\begin{aligned} X\,s = b \quad &\longrightarrow \quad s, \\ D\,w = s \quad &\longrightarrow \quad w, \\ Y\,x = w \quad &\longrightarrow \quad x, \end{aligned}$$

donde $X\,s = b$ *e* $Y\,x = w$ *se resuelven por cualquier método regresivo estable, tal como, eliminación Gaussiana con pivote parcial (GEPP) o factorización QR, y* $D\,w = s$ *es resuelto como* $w_i = s_i/d_i$, $i = 1:n$.

El Teorema 2.4.5 expresa la sensitividad de la solución del sistema $Ax = b$, donde $A = XDY$, para perturbaciones de los factores de A. Ésta depende de los números de condición de X e Y, los cuales son pequeños si $A = XDY$ es una RRD, y también de la expresión $\|A^{-1}\|\|b\|/\|x\|$, la cual es moderada para la mayoría de vectores b, según los comentarios realizados al final de la Sección 2.4.

El análisis de errores del Algoritmo 2.5.8 demuestra que éstos se pueden escribir como pequeños errores regresivos multiplicativos, y así se obtiene que:

Teorema 2.5.9 *Sea* $\|\cdot\|$ *una norma vectorial en* $\mathbb{C}^n$ *cuya norma matricial inducida satisface que* $\|\Lambda\| = \max_{i=1:n}|\lambda_i|$ *para todas las matrices diagonales* $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. *Sean* $\widehat{X}$, $\widehat{D}$, *e* $\widehat{Y}$ *los factores de* A *calculados en el* `Paso 1` *del Algoritmo 2.5.8 y supongamos que satisfacen*

$$\frac{\|\widehat{X} - X\|}{\|X\|} \leq \mathtt{u}\,p(n), \quad \frac{\|\widehat{Y} - Y\|}{\|Y\|} \leq \mathtt{u}\,p(n), \quad y \quad |\widehat{D} - D\,| \leq \mathtt{u}\,p(n)\,|D\,|, \quad i = 1:n,$$

donde X, D *e* Y *son los correspondientes factores exactos de* A, $p(n)$ *es una función poco creciente de* n, *y* `u` *es la unidad de redondeo. Supongamos también que los sistemas* $X\,s = b$ *e* $Y\,x = w$ *se resolvieron con un algoritmo regresivamente estable tal que cuando es aplicado a cualquier sistema lineal* $Bz = c$, $B \in \mathbb{C}^{n\times n}$ *y* $c \in \mathbb{C}^n$, *calcula una solución* $\widehat{z}$ *que satisface*

$$(B + \Delta B\,)\widehat{z} = c, \quad con\ \|\Delta B\,\| \leq \mathtt{u}\,q(n)\,\|B\,\|,$$

donde $q(n)$ *es una función poco creciente* n *tal que* $q(n) \geq 4\sqrt{2}/(1 - 4\mathtt{u})$. *Sea*

$$g(n) := p(n) + q(n) + \mathtt{u}\,p(n)q(n).$$

1. *Si* $\widehat{x}$ *es la solución calculada de* $A\,x = b$ *usando el Algoritmo 2.5.8, entonces*

$$(X + \Delta X\,)(D + \Delta D\,)(Y + \Delta Y\,)\widehat{x} = b,$$

donde $\|\Delta X\,\| \leq \mathtt{u}\,g(n)\,\|X\,\|$, $|\Delta D\,| \leq \mathtt{u}\,g(n)\,|D\,|$, *y* $\|\Delta Y\,\| \leq \mathtt{u}\,g(n)\,\|Y\,\|$.

2. *Además, si* x *es la solución exacta de* $Ax = b$, $(\,\mathtt{u}\,g(n)\,\kappa(Y)\,) < 1$ *y* $(\,\mathtt{u}\,g(n)\,(2+\mathtt{u}\,g(n))\,\kappa(X)\,) < 1$, *entonces*

$$\frac{\|\widehat{x} - x\|}{\|x\|} \;\leq\; \frac{\mathtt{u}\,g(n)}{1 - \mathtt{u}\,g(n)\kappa(Y)} \left(\kappa(Y) + \frac{1 + (2 + \mathtt{u}\,g(n))\,\kappa(X)}{1 - \mathtt{u}\,g(n)(2 + \mathtt{u}\,g(n))\kappa(X)} \;\frac{\|A^{-1}\|\,\|b\|}{\|x\|} \right). \tag{2.5.5}$$

Observación 2.5.10 *La ecuación* (2.5.5) *se puede simplificar considerablemente si sólo prestamos atención a los términos de primer orden*

$$\begin{aligned} \frac{\|\widehat{x} - x\|}{\|x\|} &\leq \mathtt{u}\,g(n) \left(\kappa(Y) + (1 + 2\,\kappa(X)) \;\frac{\|A^{-1}\|\,\|b\|}{\|x\|} \right) + O\left((\mathtt{u}\,g(n))^2 \right), \\ &\leq\; 4\mathtt{u}\,g(n)\;\text{máx}\{\kappa(X), \kappa(Y)\}\,\frac{\|A^{-1}\|\,\|b\|}{\|x\|} + O\left((\mathtt{u}\,g(n))^2 \right). \end{aligned}$$

Estas cotas muestran que el error relativo en la solución es $O(\mathtt{u})$ *si* $\kappa(X)$, $\kappa(Y)$ *y* $\|A^{-1}\|\,\|b\|/\|x\|$ *son pequeños.*

En los Capítulos 1 y 2 hemos presentado el estado actual del campo de la alta precisión relativa. Dentro de las ideas mostradas aquí se ha encuadrado la investigación que presentamos en esta memoria. La idea clave es la obtención de HRA mediante algoritmos que usan RRDs precisas y que luego aprovechan la estructura específica del problema. Este esquema ya se ha aplicado con éxito a problemas de valores singulares, a problemas simétricos de autovalores y a sistemas de ecuaciones lineales. Nosotros lo hemos extendido a problemas de mínimos cuadrados (Capítulo 3) y al problema simétrico de autovalores para un tipo especial de matrices: graduadas simétricas. Los trabajos existentes en el campo al comienzo de mi investigación han dado la pauta, pero en este memoria se presentan nuevos algoritmos, Algoritmo 3.5.1 y 4.3.1; nuevos análisis de errores, Teorema 3.5.2 y Corolario 4.3.5; y nuevos resultados de teoría de perturbaciones de la pseudo inversa de Moore-Penrose, Teoremas 3.2.2, 3.2.3, 3.2.5, nuevos resultados de teoría de perturbaciones de problemas de mínimos cuadrados 3.3.1 y nuevos resultados de teoría de perturbaciones de problemas de autovalores de matrices simétricas graduadas Corolario 4.2.4.

3

Teoría de perturbaciones multiplicativas y soluciones precisas del problema de mínimos cuadrados

Distintas clases de matrices con estructuras particulares aparecen frecuentemente en la teoría y las aplicaciones [58, 59]. Como consecuencia, el diseño y análisis de algoritmos especiales que involucran matrices estructuradas es un área clásica de investigación del *Álgebra Lineal Numérica* que atrae la atención de muchos investigadores. Algoritmos especiales para resolver sistemas de ecuaciones estructurados o problemas estructurados de autovalores, aparecen en muchas referencias estándar [17, 39, 43, 49, 73], pero es más raro encontrar algoritmos especiales para resolver problemas de mínimos cuadrados. En general, el objetivo, al considerar algoritmos especiales es explotar la estructura del problema para acelerar la velocidad de los cálculos y disminuir los requerimientos de memoria y mejorar la precisión de la solución en comparación con algoritmos estándar. Sobre este último objetivo, mencionemos que desde los primeros días del *Álgebra Lineal Numérica* se han desarrollado algoritmos especiales para resolver sistemas de ecuaciones lineales estructurados de manera más precisa que con algoritmos estándar (ver las referencias en [32, 43]). El desarrollo de algoritmos precisos para problemas de autovalores es mucho más reciente. Sus inicios datan del principio de los 90's con el famoso artículo [21], el cual recibió considerable atención (ver por ejemplo, [3, 20, 25, 30, 33, 35, 36, 38, 48, 66, 75] entre otras muchas referencias). El presente capítulo trata de una parte del *"Álgebra Lineal Numérica Precisa"* para la cual no existen muchas referencias disponibles en la literatura: algoritmos para resolver problemas estructurados de mínimos cuadrados $\min_x \|b - Ax\|_2$, donde $A \in \mathbb{R}^{m\times n}$ y $b \in \mathbb{R}^m$, con más precisión que la dada por algoritmos estándar y más o menos el mismo coste computacional, es decir, $O(n^2m)$ flops. Sobre este

tema sólamente conocemos la referencia [53], la cual trata de una clase particular de problemas de mínimos cuadrados. El método estándar para resolver problemas de mínimos cuadrados $\min_x \|b - Ax\|_2$ con rango completo por columnas es utilizar la factorización QR calculada con el algoritmo de Householder [43, Capítulo 19 y 20]. Dicho método es regresivamente estable, es decir, la solución calculada $\widehat{x}_0$ es la solución exacta del problema de mínimos cuadrados $\min_x \|(b+\Delta b) - (A+\Delta A)x\|_2$, donde $\|\Delta b\|_2 \leq c\,\mathtt{u}\,mn\,\|b\|_2$, $\|\Delta A\|_2 \leq c\,\mathtt{u}\,m\,n^{3/2}\,\|A\|_2$, $\mathtt{u}$ es la unidad de redondeo del ordenador y c denota una pequeña constante entera [43, Teorema 20.3]. Resultados de errores regresivos análogos se cumplen para otros métodos de solución del problema de mínimos cuadrados basado en descomposiciones ortogonales. Por ejemplo, la descomposición en valores singulares (SVD)[1]. Este resultado fuerte de errores regresivos, junto con la teoría de perturbaciones clásica "normwise" del problema de mínimos cuadrados [74, Teorema 5.1] (ver también [5, Teorema 1.4.6, p. 30]), implica la siguiente cota de error progresiva en la solución calculada $\widehat{x}_0$ con respecto a la solución exacta x_0 del problema de mínimos cuadrados

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq (c\,\mathtt{u}\,m\,n^{3/2}) \left(\kappa_2(A) + \frac{\|A^\dagger\|_2 \|b\|_2}{\|x_0\|_2} + \kappa_2(A)^2 \frac{\|b - Ax_0\|_2}{\|A\|_2\,\|x_0\|_2} \right), \quad (3.0.1)$$

donde $A^\dagger$ es la pseudoinversa de Moore-Penrose de A, $\|A\|_2$ denota la norma espectral de A y $\kappa_2(A) = \|A\|_2\,\|A^\dagger\|_2$ es el número de condición espectral de A. La cota en (3.0.1) es mayor que $\mathtt{u}\,\kappa_2(A)$ y puede ser mucho mayor en ciertas condiciones, y así, (3.0.1) no garantiza ningún dígito de precisión en la solución calculada si $\kappa_2(A) \gtrsim 1/\mathtt{u}$, es decir, si A está mal condicionada con respecto a la inversa de la unidad de redondeo. Desafortunadamente, muchos tipos de matrices estructuradas que aparecen en las aplicaciones están extremadamente mal condicionadas y los algoritmos estándar para problemas de mínimos cuadrados pueden calcular soluciones con errores relativos muy grandes. Dos ejemplos famosos son las matrices de Vandermonde, las cuales aparecen en problemas de interpolación polinomial y las matrices de Cauchy [43, Capítulos 22 y 28].

Nuestro objetivo en este trabajo es presentar un marco teórico general para la solución del problema de mínimos cuadrados y probar rigurosamente que dicho método permite calcular, para muchas clases de matrices estructuradas soluciones, con cotas de error mucho más pequeñas que la presentada en (3.0.1). El marco teórico del que

[1]Cabe señalar que el error regresivo en A cometido por la solución del problema de mínimos cuadrados vía la factorización QR de Householder es "columnwise", es decir, $\|\Delta A(:,j)\|_2 \leq c\,\mathtt{u}\,mn\,\|A(:,j)\|_2$ para $j = 1:n$ (notación de MATLAB) y, por lo tanto, este resultado es más fuerte que el mencionado anteriormente. Sin embargo, esta cota "columnwise" no se cumple para la solución vía SVD, ya que las transformaciones ortogonales son aplicadas a la matriz A por ambos lados.

hablamos se basa en el concepto de *descomposición que revela el rango* (RRD), inicialmente introducida en [20] para calcular la SVD con alta precisión relativa. nuestro método consiste en calcular la solución mínima de $\min_x \|b - Ax\|_2$ en 2 *pasos*:

1. Primer paso. Calcular una RRD precisa de $A = XDY$, en el sentido de [20] (revisamos el significado de RRD precisa en la Definición 2.4.2).

2. Segundo paso. Este paso lo realizaremos en tres etapas: (1) calcular la solución única s de $\min_x \|b - Xx\|_2$ vía la factorización QR de Householder; (2) calcular la solución w del sistema lineal $Dw = s$ como $w_i = s_i/d_i$, $i = 1 : r$; y (3) calcular la solución mínima x_0 del sistema lineal infraterminado $Yx = w$ utilizando el método Q [43, Capítulo 21]. El vector x_0 es la solución de mínima longitud de $\min_x \|b - Ax\|_2$.

Veremos que el procedimiento descrito anteriormente calcula soluciones precisas incluso para matrices A extremadamente mal condicionadas, ya que el sistema lineal mal condicionado $Dw = s$ es resuelto de manera precisa, y además que $\min_x \|b - Xx\|_2$ e $Yx = w$ también son resueltos de manera precisa ya que X e Y están bien condicionadas. Probaremos en la Sección 3.5 que el error relativo para la solución de mínima longitud $\widehat{x}_0$ de $\min_x \|b - Ax\|_2$ calculada por el método propuesto es

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \mathtt{u}\, f(m,n) \left(\kappa_2(Y) + \kappa_2(X)\, \frac{\|A^\dagger\|_2\, \|b\|_2}{\|x_0\|_2} \right), \tag{3.0.2}$$

donde $f(m,n)$ es una función moderadamente creciente de m y n. Nótese primero que (3.0.2) es mejor que (3.0.1), ya que X e Y están bien condicionadas, y así, el único factor potencialmente grande en (3.0.2) es $\|A^\dagger\|_2\, \|b\|_2/\|x_0\|_2$, el cual también aparece en (3.0.1). Pero el verdadero punto importante en la cota (3.0.2) es que si A es fija, entonces $\|A^\dagger\|_2\, \|b\|_2/\|x_0\|_2$ es pequeño para la mayoría de los vectores b, incluso para matrices A muy mal condicionadas. Este hecho es bien conocido si A es cuadrada y no singular, como lo indicamos en la Sección 2.3 (ver también [2, 14] y [32, Sección 3.2]) y, como explicaremos en la Sección 3.3.1, también se cumple para matrices generales en dos sentidos: para la mayoría de vectores aleatorios b, y para la mayoría de vectores b con un valor fijo del residuo relativo $\|Ax_0 - b\|_2/\|b\|_2$ no cercano a 1. En este capítulo la frase "para la mayoría de los vectores b" se puede entender en cualquiera de esos dos sentidos.

La idea y los resultados discutidos anteriormente se parecen a los presentados en [32] para calcular soluciones precisas de sistemas de ecuaciones lineales estructurados

$Ax = b$ con A no singular. Sin embargo, el análisis para problemas de mínimos cuadrados es mucho más complicado y requiere técnicas muy diferentes para desarrollar una nueva teoría de perturbaciones multiplicativas, necesaria para probar la cota de error presentada en (3.0.2). Además, los resultados y algoritmos que presentamos son generales: son válidos para matrices A con y sin rango completo. Aunque nos centramos principalmente en problemas de mínimos cuadrados, estos resultados pueden ser aplicados para resolver sistemas de ecuaciones lineales infradeterminados.

La mayoría de los algoritmos citados en la introducción del Capítulo 2 determinan exactamente el rango de matrices de rango deficiente y para estas clases de matrices listadas anteriormente, el método introducido en este artículo resuelve problemas de mínimos cuadrados con alta precisión relativa acotados como en (3.0.2). Esta cota de error es $O(\mathtt{u}\, f(m,n))$ para la mayoría de los vectores b independientemente del número de condición tradicional de las matrices, y así garantiza soluciones precisas.

Este capítulo está organizado como sigue. En la Sección 3.1 se introduce la notación básica, los conceptos y los resultados que se utilizarán en todo el capítulo. La Sección 3.2 estudia la variación de la pseudoinversa de Moore-Penrose bajo perturbaciones multiplicativas y, basado en estos resultados, en la Sección 3.3 presentamos cotas para perturbaciones multiplicativas de problemas de mínimos cuadrados. Como consecuencia, en la Sección 3.4 presentamos cotas perturbativas para problemas de mínimos cuadrados cuya matriz de coeficientes está dada como una RRD con factores perturbados. En la Sección 3.5 presentamos un nuevo algoritmo para calcular de manera precisa el problema de mínimos cuadrados vía la RRD y el correspondiente análisis de errores regresivo y progresivo. La precisión de este algoritmo se verifica en la Sección 3.6 por medio de exhaustivos experimentos numéricos.

3.1. Preliminares y conceptos básicos

Dado que consideramos problemas de mínimos cuadrados, usaremos la norma más natural para estos problemas: la norma vectorial euclídea, es decir, dado $x = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$, $\|x\|_2 := (|x_1|^2 + \cdots + |x_n|^2)^{1/2}$, y para matrices $A \in \mathbb{R}^{m\times n}$ la correspondiente norma matricial inducida $\|A\|_2 := \max_{\|x\|_2=1} \|Ax\|_2$, llamada la norma espectral o norma 2 de A. En la Sección 3.2.1, usaremos también normas matriciales unitariamente invariantes [69, Capítulo II. Sección 3], que serán denotadas por $\|\cdot\|$. El símbolo I_n denotará la matriz identidad de orden $n \times n$, pero usaremos simplemente I si el tamaño es claro en el contexto, y A^T denota la transpuesta de A. Usaremos la notación de MATLAB para submatrices: $A(i:j,:)$ indica la subma-

triz de A consistente de las filas i hasta la j y $A(:, k:l)$ indica la submatriz de A consistente de las columnas k hasta la l. Dada $A \in \mathbb{R}^{m\times n}$, con $m \geq n$, sus valores singulares se denotarán como $\sigma_1(A) \geq \cdots \geq \sigma_n(A) \geq 0$.

El Lema 3.1.1 será utilizado para derivar algunas cotas de perturbaciones.

Lema 3.1.1 *Sean $B, C \in \mathbb{R}^{m\times n}$, $\mathcal{S} \subseteq \mathbb{R}^m$ y $\mathcal{W} \subseteq \mathbb{R}^n$ subespacios vectoriales y sean $P_{\mathcal{S}} \in \mathbb{R}^{m\times m}$ y $P_{\mathcal{W}} \in \mathbb{R}^{n\times n}$ proyectores ortogonales sobre $\mathcal{S}$ y $\mathcal{W}$, respectivamente. Entonces se cumple:*

(a) $\|P_{\mathcal{S}}B + (I - P_{\mathcal{S}})C\|_2 \leq \sqrt{\|B\|_2^2 + \|C\|_2^2}$.

(b) $\|BP_{\mathcal{W}} + C(I - P_{\mathcal{W}})\|_2 \leq \sqrt{\|B\|_2^2 + \|C\|_2^2}$.

Demostración. (a). Sea $x \in \mathbb{R}^n$ tal que $\|x\|_2 = 1$. Dado que los vectores $P_{\mathcal{S}}Bx$ y $(I - P_{\mathcal{S}})Cx$ son ortogonales, entonces $\|(P_{\mathcal{S}}B + (I - P_{\mathcal{S}})C)x\|_2^2 = \|P_{\mathcal{S}}Bx\|_2^2 + \|(I - P_{\mathcal{S}})Cx\|_2^2 \leq \|Bx\|_2^2 + \|Cx\|_2^2 \leq \|B\|_2^2 + \|C\|_2^2$ y

$$\|P_{\mathcal{S}}B + (I - P_{\mathcal{S}})C\|_2 = \max_{\|x\|_2=1} \|(P_{\mathcal{S}}B + (I - P_{\mathcal{S}})C)x\|_2 \leq \sqrt{\|B\|_2^2 + \|C\|_2^2}.$$

La parte (b) se sigue de aplicar (a) a la transpuesta y de que para cualquier matriz, $\|A\|_2 = \|A^T\|_2$. □

Dada una matriz $G \in \mathbb{R}^{m\times n}$ con entradas g_{ij}, denotaremos por $|G|$ a la matriz con entradas $|g_{ij}|$. Expresiones como $|G| \leq |B|$, donde $B \in \mathbb{R}^{m\times n}$, significan que $|g_{ij}| \leq |b_{ij}|$ para $1 \leq i \leq m,\ 1 \leq j \leq n$.

La pseudoinversa de Moore-Penrose de $A \in \mathbb{R}^{m\times n}$ juega un papel importante en este trabajo. Es bien sabido que la única matriz $Z \in \mathbb{R}^{n\times m}$ tal que

$$\text{(i) } AZA = A, \quad \text{(ii) } ZAZ = Z, \quad \text{(iii) } (AZ)^T = AZ, \quad \text{y} \quad \text{(iv) } (ZA)^T = ZA, \tag{3.1.1}$$

o equivalentemente, tal que

$$AZ = P_A \quad \text{y} \quad ZA = P_Z, \tag{3.1.2}$$

donde P_A y P_Z son los proyectores ortogonales sobre el espacio columna de A y de Z, respectivamente. La equivalencia de las cuatro condiciones en (3.1.1) y las dos condiciones en (3.1.2) pueden ser encontradas en [10, Teorema 1.1.1]. Denotaremos por $A^\dagger \in \mathbb{R}^{m\times n}$ la pseudoinversa de Moore-Penrose de $A \in \mathbb{R}^{m\times n}$. Recordemos

que si $A \in \mathbb{R}^{n\times n}$ es no singular, entonces $A^\dagger = A^{-1}$ y también que la SVD de A nos permite obtener una expresión para $A^\dagger$ y probar muchas de sus propiedades [69, Capítulo 3]. $\mathcal{R}(A)$ y $\mathcal{N}(A)$ denotarán respectivamente el espacio columna y el espacio nulo de A. Es fácil ver que $\mathcal{R}(A^T) = \mathcal{R}(A^\dagger)$, entonces, por (3.1.2), $P_A = AA^\dagger$ y $P_{A^T} = P_{A^\dagger} = A^\dagger A$ son respectivamente los proyectores ortogonales sobre $\mathcal{R}(A)$ y $\mathcal{R}(A^T)$.

Enunciamos en el Lema 3.1.2 algunas propiedades muy conocidas [10] de la pseudoinversa de Moore-Penrose que necesitaremos a lo largo de este capítulo.

Lema 3.1.2 (a) *Si A tiene rango completo por filas, entonces $A^\dagger = A^T(AA^T)^{-1}$ y $AA^\dagger = I$.*

(b) *Si A tiene rango completo por columnas, entonces $A^\dagger = (A^TA)^{-1}A^T$ y $A^\dagger A = I$.*

(c) *Sean $F \in \mathbb{R}^{m\times r}$ y $G \in \mathbb{R}^{r\times n}$. Si rango (F) = rango $(G) = r$, entonces $(FG)^\dagger = G^\dagger F^\dagger$.*

La solución de mínima longitud del problema de mínimos cuadrados $\min_{x\in\mathbb{R}^n} \|b - Ax\|_2$ es $x_0 = A^\dagger\, b$ y la solución de mínima longitud de un sistema lineal indeterminado $Ax = b$ es también $x_0 = A^\dagger\, b$. Si $A = XDY \in \mathbb{R}^{m\times n}$ es una RRD de A, entonces dos aplicaciones del Lema 3.1.2-(c) implican que $A^\dagger = Y^\dagger D^{-1} X^\dagger$ y la solución de mínima longitud del problema de mínimos cuadrados o de un sistema lineal indeterminado es $x_0 = Y^\dagger D^{-1} X^\dagger\, b$.

3.2. Perturbaciones multiplicativas para la pseudoinversa de Moore-Penrose

En esta sección y en la Sección 3.3, consideramos una perturbación multiplicativa de una matriz general $A \in \mathbb{R}^{m\times n}$, es decir, una matriz $\widetilde{A} = (I + E)A(I + F)$, donde $(I + E) \in \mathbb{R}^{m\times m}$ e $(I + F) \in \mathbb{R}^{n\times n}$ son matrices no singulares. El objetivo final es acotar en la Sección 3.3, la expresión $\|\widetilde{x}_0 - x_0\|_2/\|x_0\|_2$, donde x_0 y $\widetilde{x}_0$ son las soluciones de mínima longitud de los problemas de mínimos cuadrados $\min_{x\in\mathbb{R}^n} \|Ax - b\|_2$ y $\min_{x\in\mathbb{R}^n} \|\widetilde{A}x - \widetilde{b}\|_2$, respectivamente. Este objetivo se alcanza a través del teorema principal de esta sección, el Teorema 3.2.2, donde obtenemos dos expresiones para $\widetilde{A}^\dagger$ en terminos de $A^\dagger$, $(I + E)^{-1}$ e $(I + F)^{-1}$. Utilizamos las expresiones de $\widetilde{A}^\dagger$ para

desarollar en la Sección 3.2.1 cotas para $\|\widetilde{A}^{\dagger} - A^{\dagger}\|/\|A^{\dagger}\|$ en cualquier norma unitariamente invariante y en particular, en la norma 2. Aunque estas cotas no son necesarias para nuestro resultado final, resaltamos que encontrar cotas perturbativas para la pseudoinversa de Moore-Penrose es un tema clásico en el Análisis Matricial (ver [74] y [69, Capítulo 3, Sección 3]) que ha llamado la atención de muchos investigadores. Mostramos que los resultados en la Sección 3.2.1 son mejores que los presentados en [9], los cuales tienen una naturaleza diferente y son obtenidos a través de un método diferente. La teoría multiplicativa de perturbaciones de matrices es muy importante en el cálculo de autovalores y valores singulares con alta precisión [37, 45, 46, 51, 52] y también el cálculo de soluciones precisas de sistemas de ecuaciones [32, Lema 3.1]. Hasta donde sabemos, no ha sido estudiada en el contexto de encontrar soluciones precisas del problema de mínimos cuadrados.

El Lema 3.2.1 es un resultado técnico que utilizaremos en la demostración del Teorema 3.2.2.

Lema 3.2.1 *Sean $A \in \mathbb{R}^{m\times n}$ y $\widetilde{A} = (I+E)A(I+F) \in \mathbb{R}^{m\times n}$, donde $(I+E) \in \mathbb{R}^{m\times m}$ e $(I+F) \in \mathbb{R}^{n\times n}$ son matrices no singulares. Entonces las siguientes igualdades se cumplen:*

(a) $P_A(I+E^T)(I-P_{\widetilde{A}}) = 0.$

(b) $(I-P_{\widetilde{A}^T})(I+F^T)P_{A^T} = 0.$

Demostración. (a) Dado que $\mathcal{R}(\widetilde{A}) = \mathcal{R}((I+E)A)$ entonces $(I-P_{\widetilde{A}})(I+E)A = 0$. Así que, $(I-P_{\widetilde{A}})(I+E)AA^{\dagger} = (I-P_{\widetilde{A}})(I+E)P_A = 0$, lo cual es equivalente a $P_A(I+E^T)(I-P_{\widetilde{A}}) = 0$.

(b) Aplicar la parte (a) a $\widetilde{A}^T = (I+F^T)A^T(I+E^T)$, luego conjugar y transponer la igualdad. □

Ahora, enunciamos el resultado principal de esta sección, el cual es válido para matrices de rango completo, matrices de rango deficiente y para perturbaciones de cualquier magnitud.

Teorema 3.2.2 *Sean $A \in \mathbb{R}^{m\times n}$ y $\widetilde{A} = (I+E)A(I+F) \in \mathbb{R}^{m\times n}$, donde $(I+E) \in \mathbb{R}^{m\times m}$ e $(I+F) \in \mathbb{R}^{n\times n}$ son matrices no singulares. Entonces*

$$\widetilde{A}^{\dagger} = P_{\widetilde{A}^T}(I+F)^{-1}A^{\dagger}(I+E)^{-1}P_{\widetilde{A}} \tag{3.2.1}$$

y

$$\widetilde{A}^{\dagger} = \left(I + (I - P_{\widetilde{A}^T})F^T - P_{\widetilde{A}^T}\widehat{F}\right) A^{\dagger} \left(I + E^T(I - P_{\widetilde{A}}) - \widehat{E}P_{\widetilde{A}}\right), \tag{3.2.2}$$

donde $\widehat{E} = (I + E)^{-1}E$ *y* $\widehat{F} = (I + F)^{-1}F$.

Demostración. Definamos $Z := P_{\widetilde{A}^T}(I + F)^{-1}A^{\dagger}(I + E)^{-1}P_{\widetilde{A}}$. Probaremos que Z satisface las condiciones (3.1.2) reemplazando A por $\widetilde{A}$. Recordemos que $P_{\widetilde{A}^T} = \widetilde{A}^{\dagger}\widetilde{A}$ y $P_{\widetilde{A}} = \widetilde{A}\widetilde{A}^{\dagger}$. Entonces

$$\begin{aligned}\widetilde{A}Z &= \widetilde{A}(I + F)^{-1}A^{\dagger}(I + E)^{-1}P_{\widetilde{A}} = (I + E)AA^{\dagger}(I + E)^{-1}P_{\widetilde{A}} \\ &= (I + E)AA^{\dagger}(I + E)^{-1}\widetilde{A}\widetilde{A}^{\dagger} = (I + E)AA^{\dagger}A(I + F)\widetilde{A}^{\dagger} = \widetilde{A}\widetilde{A}^{\dagger} = P_{\widetilde{A}}.\end{aligned}$$

Análogamente,

$$\begin{aligned}Z\widetilde{A} &= P_{\widetilde{A}^T}(I + F)^{-1}A^{\dagger}(I + E)^{-1}\widetilde{A} = P_{\widetilde{A}^T}(I + F)^{-1}A^{\dagger}A(I + F) \\ &= \widetilde{A}^{\dagger}\widetilde{A}(I + F)^{-1}A^{\dagger}A(I + F) = \widetilde{A}^{\dagger}(I + E)AA^{\dagger}A(I + F) = \widetilde{A}^{\dagger}\widetilde{A} = P_{\widetilde{A}^T}.\end{aligned} \tag{3.2.3}$$

La igualdad (3.2.3) implica que $\mathcal{R}(\widetilde{A}^T) \subseteq \mathcal{R}(Z)$ y la definición de Z implica que $\mathcal{R}(Z) \subseteq \mathcal{R}(\widetilde{A}^T)$. Así, $\mathcal{R}(Z) = \mathcal{R}(\widetilde{A}^T)$ y la ecuación (3.2.3) implica $Z\widetilde{A} = P_Z$. Por lo tanto, las condiciones (3.1.2) se cumplen para $\widetilde{A}$ y $Z = \widetilde{A}^{\dagger}$, con lo que se demuestra (3.2.1).

A continuación, utilizamos (3.2.1) para demsotrar (3.2.2). Primero, escribimos $(I + E)^{-1} = I - (I + E)^{-1}E = I - \widehat{E}$ e $(I + F)^{-1} = I - (I + F)^{-1}F = I - \widehat{F}$. Sustituyendo estas expresiones en (3.2.1), tenemos

$$\widetilde{A}^{\dagger} = P_{\widetilde{A}^T}(I - \widehat{F})A^{\dagger}(I - \widehat{E})P_{\widetilde{A}} = P_{\widetilde{A}^T}(P_{A^T} - \widehat{F})A^{\dagger}(P_A - \widehat{E})P_{\widetilde{A}}. \tag{3.2.4}$$

Por el Lema 3.2.1-(a) se sigue que $P_A(I + E^T(I - P_{\widetilde{A}})) = P_AP_{\widetilde{A}}$. Análogamente, del Lema 3.2.1-(b) tenemos que, $((I - P_{\widetilde{A}^T})F^T + I)P_{A^T} = P_{\widetilde{A}^T}P_{A^T}$. Finalmente, sustituyendo estas relaciones en (3.2.4), y usando que $A^{\dagger}P_A = A^{\dagger}$ y $P_{A^T}A^{\dagger} = A^{\dagger}$, obtenemos (3.2.2). □

Si $m = n$ y A es no singular, entonces $P_{\widetilde{A}} = P_{\widetilde{A}^T} = I_n$, (3.2.1) y (3.2.2) se convierten en $\widetilde{A}^{-1} = (I+F)^{-1}A^{-1}(I+E)^{-1}$. Si A tiene rango completo por columnas, entonces $\widetilde{A}$ tiene rango completo por columnas, $P_{\widetilde{A}^T} = I_n$, (3.2.1) se simplifica a $\widetilde{A}^{\dagger} = (I + F)^{-1}A^{\dagger}(I + E)^{-1}P_{\widetilde{A}}$ y (3.2.2) a $\widetilde{A}^{\dagger} = \left(I - \widehat{F}\right) A^{\dagger} \left(I + E^T(I - P_{\widetilde{A}}) - \widehat{E}P_{\widetilde{A}}\right)$. Finalmente, si A tiene rango completo por filas, entonces $\widetilde{A}$ también tiene rango

completo por filas, $P_{\widetilde{A}} = I_m$, (3.2.1) se simplifica a $\widetilde{A}^\dagger = P_{\widetilde{A}^T}(I+F)^{-1}A^\dagger(I+E)^{-1}$ y (3.2.2) a $\widetilde{A}^\dagger = \left(I + (I - P_{\widetilde{A}^T})F^T - P_{\widetilde{A}^T}\widehat{F}\right)A^\dagger\left(I - \widehat{E}\right)$.

Detacamos que la expresión (3.2.2) garantiza, que bajo pequeñas perturbaciones multiplicativas de A, obtenemos pequeñas perturbaciones multiplicativas de $A^\dagger$.

Las hipótesis del Teorema 3.2.2 garantizan que $rango\,(A) = rango\,(\widetilde{A})$. Esto simplifica considerablemente el análisis del cambio de la pseudoinversa de Moore-Penrose con respecto a perturbaciones aditivas $\widetilde{A} = A + \Delta A$ [69, 74]. Además, el Teorema 3.2.3 implica que si la condición débil máx$\{\|E\|_2, \|F\|_2\} < 1$ se cumple, entonces $\widetilde{A} = (I+E)A(I+F)$ es una *perturbación aguda* de A (ver la definición en [74, Definición 7.2] y también en [69, Capítulo III, Definición 3.2]). Es bien conocido que *perturbaciones agudas* introducen simplificaciones aún para perturbaciones aditivas $\widetilde{A} = A + \Delta A$. Las cotas en el Teorema 3.2.3 serán utilizadas en la Sección 3.3.

Teorema 3.2.3 *Sean $A \in \mathbb{R}^{m\times n}$ y $\widetilde{A} = (I+E)A(I+F) \in \mathbb{R}^{m\times n}$, donde $(I+E) \in \mathbb{R}^{m\times m}$ e $(I+F) \in \mathbb{R}^{n\times n}$ son matrices no singulares, además consideremos $P_{\mathcal{N}(A)}$ y $P_{\mathcal{N}(\widetilde{A})}$ los proyectores ortogonales sobre el espacio nulo de A y $\widetilde{A}$, respectivamente. Entonces:*

(a) $\|P_{\widetilde{A}} - P_A\|_2 = \|P_{\widetilde{A}}(I - P_A)\|_2 = \|P_A(I - P_{\widetilde{A}})\|_2 \leq \|E\|_2$.

(b) $\|P_{\widetilde{A}^T} - P_{A^T}\|_2 = \|P_{\widetilde{A}^T}(I - P_{A^T})\|_2 = \|P_{A^T}(I - P_{\widetilde{A}^T})\|_2 \leq \|F\|_2$.

(c) $\|P_{\mathcal{N}(\widetilde{A})} - P_{\mathcal{N}(A)}\|_2 \leq \|F\|_2$.

Demostración. (a). Los subespacios $\mathcal{R}(A)$ y $\mathcal{R}(\widetilde{A})$ tienen la misma dimensión. Así que, de [69, Capítulo I, Teorema 5.5], $\|P_{\widetilde{A}} - P_A\|_2 = \|P_{\widetilde{A}}(I - P_A)\|_2 = \|P_A(I - P_{\widetilde{A}})\|_2$. Además, por el Lema 3.2.1-(a), $\|P_A(I - P_{\widetilde{A}})\|_2 = \| - P_A E^T(I - P_{\widetilde{A}})\|_2 \leq \|E^T\|_2 = \|E\|_2$.

La parte (b) se demuestra aplicando (a) a la expresión $\widetilde{A}^T = (I+F^T)A^T(I+E^T)$. Finalmente, la parte (c) se obtiene de (b), $P_{\mathcal{N}(A)} = I - P_{A^T}$ y $P_{\mathcal{N}(\widetilde{A})} = I - P_{\widetilde{A}^T}$. □

El Corolario 3.2.4 proporciona una expresión para $\widetilde{A}^\dagger - A^\dagger$ que se obtiene directamente de (3.2.2). Este corolario será utilizado en la Sección 3.2.1 y en el Teorema 3.3.1.

Corolario 3.2.4 *Sean $A \in \mathbb{R}^{m\times n}$ y $\widetilde{A} = (I+E)A(I+F) \in \mathbb{R}^{m\times n}$, donde $(I+E) \in \mathbb{R}^{m\times m}$ e $(I+F) \in \mathbb{R}^{n\times n}$ son matrices no singulares, $\widehat{E} = (I+E)^{-1}E$ y $\widehat{F} = (I+F)^{-1}F$. Entonces*

$$\widetilde{A}^\dagger - A^\dagger = A^\dagger\Theta_E + \Theta_F A^\dagger + \Theta_F A^\dagger \Theta_E, \tag{3.2.5}$$

donde

$$\Theta_E = E^T(I - P_{\widetilde{A}}) - \widehat{E}P_{\widetilde{A}} \quad y \quad \Theta_F = (I - P_{\widetilde{A}^T})F^T - P_{\widetilde{A}^T}\widehat{F}\,. \tag{3.2.6}$$

3.2.1. Cotas de perturbaciones multiplicativas para la pseudoinversa de Moore-Penrose

Como explicamos en el primer párrafo de la Sección 3.2, los resultados en esta sección no serán utilizados en ningún otro lugar del capítulo. El objetivo principal en esta sección es presentar cotas para $\|\widetilde{A}^\dagger - A^\dagger\|/\|A^\dagger\|$. Supongamos que $A \neq 0$, ya que en otro caso el problema es trivial. Como resultado principal tenemos el Teorema 3.2.5.

Teorema 3.2.5 *Sean $A \in \mathbb{R}^{m\times n}$ y $\widetilde{A} = (I+E)A(I+F) \in \mathbb{R}^{m\times n}$, donde $(I+E) \in \mathbb{R}^{m\times m}$ e $(I+F) \in \mathbb{R}^{n\times n}$ son matrices no singulares, y sean $\widehat{E} = (I+E)^{-1}E$ y $\widehat{F} = (I+F)^{-1}F$. Denotemos por $\|\cdot\|$ una norma invariante unitariamente normalizada y por $\|\cdot\|_2$ la norma espectral. Entonces:*

(a) $$\frac{\|\widetilde{A}^\dagger - A^\dagger\|}{\min\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}} \leq \|E\| + \|\widehat{E}\| + \|F\| + \|\widehat{F}\| + \left(\|E\| + \|\widehat{E}\|\right)\left(\|F\| + \|\widehat{F}\|\right).$$

(b) $$\frac{\|\widetilde{A}^\dagger - A^\dagger\|_2}{\|A^\dagger\|_2} \leq \sqrt{\|E\|_2^2 + \|F\|_2^2 + \left(\|\widehat{E}\|_2 + \|\widehat{F}\|_2 + \|\widehat{E}\|_2\,\|\widehat{F}\|_2\right)^2} \quad y$$

$$\frac{\|\widetilde{A}^\dagger - A^\dagger\|_2}{\|\widetilde{A}^\dagger\|_2} \leq \sqrt{\|\widehat{E}\|_2^2 + \|\widehat{F}\|_2^2 + \left(\|E\|_2 + \|F\|_2 + \|E\|_2\,\|F\|_2\right)^2}\,.$$

Demostración. (a). La cota para $\|\widetilde{A}^\dagger - A^\dagger\|/\|A^\dagger\|_2$ se obtiene directamente de (3.2.5), teniendo sólamente en cuenta que para cualquier matriz B y C, $\|BC\| \leq \|B\|_2\|C\|$ y $\|BC\| \leq \|B\|\|C\|_2$ [69, página 80]. La cota para $\|\widetilde{A}^\dagger - A^\dagger\|/\|\widetilde{A}^\dagger\|_2$ se obtiene de la cota para $\|\widetilde{A}^\dagger - A^\dagger\|/\|A^\dagger\|_2$ intercambiando los roles de A y $\widetilde{A}$, es decir, considerando A como una perturbación multiplicativa $\widetilde{A}$, o bien $A = (I+E)^{-1}\widetilde{A}(I+F)^{-1} = (I-\widehat{E})\widetilde{A}(I-\widehat{F})$, esto equivale a intercambiar $\|E\|$ y $\|F\|$ en las cotas por $\|\widehat{E}\|$ y $\|\widehat{F}\|$, respectivamente, y vice versa.

(b). Observemos que de (3.2.6), el Lema 3.2.1-(b) y $P_{A^T} = A^\dagger A$, tenemos

$$\begin{aligned}(I+\Theta_F)A^\dagger A &= (I+(I-P_{\widetilde{A}^T})F^T - P_{\widetilde{A}^T}\widehat{F})P_{A^T}\\ &= P_{A^T} + (I-P_{\widetilde{A}^T})F^T P_{A^T} - P_{\widetilde{A}^T}\widehat{F}P_{A^T}\\ &= P_{A^T} - (I-P_{\widetilde{A}^T})P_{A^T} - P_{\widetilde{A}^T}\widehat{F}P_{A^T}\\ &= P_{\widetilde{A}^T}(I-\widehat{F})P_{A^T}\,,\end{aligned}$$

y si multiplicamos por la derecha la expresión anterior por $A^\dagger$

$$(I+\Theta_F)A^\dagger = P_{\widetilde{A}^T}(I-\widehat{F})A^\dagger\,. \tag{3.2.7}$$

Similarmente, pero usando el Lema 3.2.1-(a), tenemos

$$A^\dagger\,(I+\Theta_E) = A^\dagger(I-\widehat{E})P_{\widetilde{A}}\,. \tag{3.2.8}$$

Ahora, usamos (3.2.5), (3.2.7)-(3.2.8) y (3.2.6) para demostrar que

$$\begin{aligned}\widetilde{A}^\dagger - A^\dagger &= (I+\Theta_F)A^\dagger\Theta_E + \Theta_F A^\dagger\\ &= P_{\widetilde{A}^T}(I-\widehat{F})A^\dagger\Theta_E + \Theta_F A^\dagger\\ &= P_{\widetilde{A}^T}\left[(I-\widehat{F})A^\dagger\Theta_E - \widehat{F}A^\dagger\right] + (I-P_{\widetilde{A}^T})F^T A^\dagger\\ &= P_{\widetilde{A}^T}\left[A^\dagger\Theta_E - \widehat{F}A^\dagger(I+\Theta_E)\right] + (I-P_{\widetilde{A}^T})F^T A^\dagger\\ &= P_{\widetilde{A}^T}\left[A^\dagger E^T(I-P_{\widetilde{A}}) - A^\dagger\widehat{E}P_{\widetilde{A}} - \widehat{F}A^\dagger(I-\widehat{E})P_{\widetilde{A}}\right] + (I-P_{\widetilde{A}^T})F^T A^\dagger\,.\end{aligned} \tag{3.2.9}$$

Queda entonces aplicar el Lema 3.1.1 a la ecuación (3.2.9) y obtenemos

$$\begin{aligned}\|\widetilde{A}^\dagger - A^\dagger\|_2^2 &\le \|A^\dagger E^T(I-P_{\widetilde{A}}) - [A^\dagger\widehat{E} + \widehat{F}A^\dagger(I-\widehat{E})]\,P_{\widetilde{A}}\|_2^2 + \|F^T A^\dagger\|_2^2\\ &\le \|A^\dagger E^T\|_2^2 + \|A^\dagger\widehat{E} + \widehat{F}A^\dagger(I-\widehat{E})\|_2^2 + \|F\|_2^2\,\|A^\dagger\|_2^2\\ &\le \|A^\dagger\|_2^2\left(\|E\|_2^2 + \|F\|_2^2 + (\|\widehat{E}\|_2 + \|\widehat{F}\|_2 + \|\widehat{E}\|_2\,\|\widehat{F}\|_2)^2\right)\,,\end{aligned}$$

lo cual nos proporciona una cota para $\|\widetilde{A}^\dagger - A^\dagger\|_2/\|A^\dagger\|_2$ en la parte (b). Con lo que, la cota para $\|\widetilde{A}^\dagger - A^\dagger\|_2/\|\widetilde{A}^\dagger\|_2$ se obtiene intercambiando los roles de A y $\widetilde{A}$ como lo hicimos en la demostración de la parte (a). □

Observación 3.2.6 Resaltamos los siguientes puntos del Teorema 3.2.5.

(a) Las cotas en el Teorema 3.2.5 mejoran significativamente las cotas clásicas para el cambio relativo de la pseudoinversa de Moore-Penrose a perturbaciones aditivas $\widetilde{A} = A + \Delta A$ (ver [74, Teorema 4.1] y [69, Capítulo III, Corolario 3.10]). El punto importante es que las cotas en el Teorema 3.2.5 no dependen de $\kappa_2(A)$, mientras que las cotas clásicas si.

(b) La cota en la parte (a) del Teorema 3.2.5 tiene la ventaja de ser válida para cualquier norma invariante unitariamente normalizada, pero para la norma 2, la cota en la parte (b) es siempre mejor que la dada en la parte (a), dado que $\sqrt{x^2+y^2} \leq x+y$, para números reales x e y tales que $x \geq 0$, $y \geq 0$ y

$$\sqrt{\|E\|_2^2 + \|F\|_2^2 + \left(\|\widehat{E}\|_2 + \|\widehat{F}\|_2 + \|\widehat{E}\|_2 \, \|\widehat{F}\|_2\right)^2} \leq$$

$$\leq \sqrt{\|E\|_2^2 + \|F\|_2^2} \; + \; \|\widehat{E}\|_2 + \|\widehat{F}\|_2 + \|\widehat{E}\|_2 \, \|\widehat{F}\|_2$$
$$\leq \|E\|_2 + \|\widehat{E}\|_2 + \|F\|_2 + \|\widehat{F}\|_2 + (\|E\|_2 + \|\widehat{E}\|_2)\,(\|F\|_2 + \|\widehat{F}\|_2)\,.$$

(c) Si A tiene rango completo por filas, entonces $\widetilde{A}$ tiene rango completo por filas y además $P_{\widetilde{A}} = I_m$. Así que, la expresión para Θ_E que aparece en (3.2.6) se simplifica a $\Theta_E = -\widehat{E}$ y todos los términos que contienen $\|E\|$ ó $\|E\|_2$ en las cotas del Teorema 3.2.5 se anulan (pero es necesario mantener $\|\widehat{E}\|$ y $\|\widehat{E}\|_2$).

(d) Si A tiene rango completo por columnas, entonces $\widetilde{A}$ tiene rango completo por columnas y además $P_{\widetilde{A}^T} = I_n$. Así que, la expresión para Θ_F que aparece en (3.2.6) se simplifica a $\Theta_F = -\widehat{F}$ y todos los términos que contienen $\|F\|$ ó $\|F\|_2$ en las cotas del Teorema 3.2.5 se anulan (pero es necesario mantener $\|\widehat{F}\|$ y $\|\widehat{F}\|_2$).

(e) Si en el Teorema 3.2.5 restringimos el tamaño de las perturbaciones a máx$\{\|E\|_2, \|F\|_2\} < 1$, una condición que garantiza que $(I+E)$ e $(I+F)$ sean no singulares, entonces las desigualdades estándar de normas matriciales [39] implican

$$\|\widehat{E}\|_2 \leq \frac{\|E\|_2}{1-\|E\|_2} \quad \text{y} \quad \|\widehat{F}\|_2 \leq \frac{\|F\|_2}{1-\|F\|_2}\,. \tag{3.2.10}$$

Estas desigualdades pueden ser usadas en la parte (b) del Teorema 3.2.5 para obtener cotas en términos de $\|E\|_2$ y $\|F\|_2$.

(g) Finalmente, con la restricción adicional máx$\{\|E\|_2, \|F\|_2\} < 1$, el Teorema 3.2.5 puede completarse con

$$\frac{\|A^\dagger\|_2}{(1+\|E\|_2)(1+\|F\|_2)} \leq \|\widetilde{A}^\dagger\|_2 \leq \frac{\|A^\dagger\|_2}{(1-\|E\|_2)(1-\|F\|_2)}.$$

El lado derecho de la desigualdad proviene de (3.2.1), lo cual implica que $\|\widetilde{A}^\dagger\|_2 \leq \|(I+F)^{-1}\|_2 \, \|A^\dagger\|_2 \, \|(I+E)^{-1}\|_2$. Para la desigualdad del lado izquierdo: consideremos A como una perturbación multiplicativa de $\widetilde{A}$, es decir,

$A = (I+E)^{-1}\widetilde{A}(I+F)^{-1}$, y apliquemos (3.2.1) con los roles de A y $\widetilde{A}$ intercambiados para obtener $A^\dagger = P_{A^T}(I+F)\widetilde{A}^\dagger(I+E)P_A$. Esto implica que $\|A^\dagger\|_2 \leq (1+\|F\|_2)\|\widetilde{A}^\dagger\|_2(1+\|E\|_2)$.

Recientemente en [9, Sección 4] fueron presentadas cotas de perturbaciones multiplicativas para la pseudoinversa de Moore-Penrose. Las cotas presentadas en [9] no están basadas en expresiones para $\widetilde{A}^\dagger$ como las del Teorema 3.2.2, son obtenidas siguiendo un método diferente. En el resto de esta sección comparamos las cotas presentadas en el Teorema 3.2.5 con las propuestas en [9]. Escribimos las cotas multiplicativas [2] dadas en [9, Teoremas 4.1 y 4.2] con la misma notación del Teorema 3.2.5:

$$\frac{\|\widetilde{A}^\dagger - A^\dagger\|}{\text{máx}\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}} \leq \|E\| + \|\widehat{E}\| + \|F\| + \|\widehat{F}\|\,, \tag{3.2.11}$$

$$\frac{\|\widetilde{A}^\dagger - A^\dagger\|_2}{\text{máx}\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}} \leq \sqrt{\frac{3}{2}}\,\sqrt{\|E\|_2^2 + \|\widehat{E}\|_2^2 + \|F\|_2^2 + \|\widehat{F}\|_2^2}\,. \tag{3.2.12}$$

Notemos que la presencia del $\text{máx}\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}$ dificulta comparar las cotas presentadas en (3.2.11)-(3.2.12) con las cotas presentadas en el Teorema 3.2.5. Por ejemplo, si $\|A^\dagger\|_2 = \|\widetilde{A}^\dagger\|_2$, entonces la cota en (3.2.11) es obviamente mejor que la del Teorema 3.2.5-(a), pero si $\|A^\dagger\|_2 \ll \|\widetilde{A}^\dagger\|_2$, entonces (3.2.11) no da información sobre $\|\widetilde{A}^\dagger - A^\dagger\|/\|A^\dagger\|_2$, mientras que el Teorema 3.2.5-(a) si. *Sin embargo, como discutiremos a continuación, las cotas en el Teorema 3.2.5 son mejores que las presentadas en* (3.2.11)-(3.2.12) *ambas a primer orden.*

Si consideramos perturbaciones muy pequeñas e ignoramos los teŕminos de segundo orden, entonces podemos reemplazar $\text{máx}\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}$ y $\text{mín}\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}$, simplemente por $\|A^\dagger\|_2$, esto permite hacer comparaciones fácilmente. El Teorema 3.2.5-(a) y la ecuación (3.2.11) proporcionan la misma cota a primer orden, es decir, $\|\widetilde{A}^\dagger - A^\dagger\|/\|A^\dagger\|_2 \leq 2(\|E\| + \|F\|)$. Sin embargo, a primer orden, el lado derecho de la ecuación (3.2.12) es $\Xi = \sqrt{3}\,\sqrt{\|E\|_2^2 + \|F\|_2^2}$ y la cota en el Teorema 3.2.5-(b) es $\Gamma_b = \sqrt{\|E\|_2^2 + \|F\|_2^2 + (\|E\|_2 + \|F\|_2)^2}$. Para comparar Γ_b y Ξ, utilizamos que $(x+y)^2 \leq 2(x^2+y^2)$, para números reales x e y tales que $x \geq 0$ e $y \geq 0$. Así que

$$\Gamma_b \leq \sqrt{3\,\|E\|_2^2 + 3\,\|F\|_2^2} = \Xi,$$

implica, que a primer orden, la cota en el Teorema 3.2.5-(b) es siempre mejor que la presentada en (3.2.12).

[2] la cota dada en (3.2.12) es presentada en [9] para la clase de normas unitariamente invariantes llamadas *Q-normas*, las cuales incluyen, entre otras, la norma espectral y la de Frobenius.

Para perturbaciones suficientemente grandes, la presencia del máx$\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}$ hace que las ecuaciones (3.2.11) y (3.2.12) sean inaplicables en ciertas situaciones, ya que uno de los objetivos estándar de la teoría de perturbaciones es acotar $\|\widetilde{A}^\dagger - A^\dagger\|$ *sin conocer* $\widetilde{A}^\dagger$ *y teniendo sólamente algunas cotas sobre las normas de las perturbaciones* E *y* F. Ilustramos esto con un ejemplo. Sean $A = [1\ 0; 0\ 1; 0\ 0] \in \mathbb{R}^{3\times 2}$ (hemos utilizado la notación de MATLAB para las matrices), $E = diag(-4/5, -4/5, -4/5)$ y $F = diag(-4/5, -4/5)$. Un cálculo sencillo muestra que

$$\frac{\|\widetilde{A}^\dagger - A^\dagger\|_2}{\|A^\dagger\|_2} = 24, \quad \frac{\|\widetilde{A}^\dagger - A^\dagger\|_2}{\|\widetilde{A}^\dagger\|_2} = 0{,}96$$
$$\|E\|_2 = 0{,}8, \quad \|F\|_2 = 0{,}8, \quad \|\widehat{E}\|_2 = 4, \quad \|\widehat{F}\|_2 = 4,$$

los cuales dan: 9,6 para la cota en (3.2.11); 32,64 para la cota en el Teorema 3.2.5-(a); 7,07 para la cota en (3.2.12); 24,03 para la primera cota del Teorema 3.2.5-(b); y 6,084 para la segunda cota del Teorema 3.2.5-(b). Por lo tanto, en este ejemplo, las ecuaciones (3.2.11)-(3.2.12) fallan al dar una cota para $\|\widetilde{A}^\dagger - A^\dagger\|_2/\|A^\dagger\|_2$, mientras que las cotas en el Teorema 3.2.5 dan estimaciones fuertes (en particular aquellas en el Teorema 3.2.5-(b)).

3.3. Perturbaciones multiplicativas para problemas de mínimos cuadrados

En esta sección consideramos el problema de mínimos cuadrados

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2, \quad A \in \mathbb{R}^{m\times n}, \quad b \in \mathbb{R}^m, \tag{3.3.1}$$

y el problema de mínimos cuadrados perturbado

$$\min_{x\in\mathbb{R}^n} \|\widetilde{A}x - \widetilde{b}\|_2, \quad \widetilde{A} = (I+E)A(I+F) \in \mathbb{R}^{m\times n}, \ \ \widetilde{b} = b + h \in \mathbb{R}^m\,, \tag{3.3.2}$$

donde $(I + E) \in \mathbb{R}^{m\times m}$ e $(I + F) \in \mathbb{R}^{n\times n}$ son matrices no singulares. Nuestro objetivo es encontrar una cota superior para el cambio relativo $\|\widetilde{x}_0 - x_0\|_2/\|x_0\|_2$, donde $x_0 = A^\dagger b$ y $\widetilde{x}_0 = \widetilde{A}^\dagger\widetilde{b}$ son respectivamente las soluciones de mínima longitud de (3.3.1) y (3.3.2). También examinamos el cambio de los residuos asociados $r = b - Ax_0$ y $\widetilde{r} = \widetilde{b} - \widetilde{A}\widetilde{x}_0$. El Teorema 3.3.1 es el resultado principal en esta sección.

Teorema 3.3.1 *Sean* x_0 *y* $\widetilde{x}_0$ *respectivamente las soluciones de mínima longitud de (3.3.1) y (3.3.2), y sean* $r = b - Ax_0$ *y* $\widetilde{r} = \widetilde{b} - \widetilde{A}\widetilde{x}_0$ *los correspondientes residuos.*

Sea $\widehat{E} = (I+E)^{-1}E$ *y* $\widehat{F} = (I+F)^{-1}F$, *definamos* $\alpha_E := \sqrt{\|E\|_2^2 + \|\widehat{E}\|_2^2}$ *y* $\alpha_F := \sqrt{\|F\|_2^2 + \|\widehat{F}\|_2^2}$, *y asumamos que* $\|h\|_2 \leq \epsilon\|b\|_2$. *Entonces las siguientes cotas se cumplen*

$$\|\widetilde{x}_0 - x_0\|_2 \leq \alpha_F \|x_0\|_2 + [\alpha_E (1+\alpha_F)(1+\epsilon) + \epsilon(1+\alpha_F)] \|A^\dagger\|_2 \|b\|_2 \, y \tag{3.3.3}$$

$$\|\widetilde{r} - r\|_2 \leq \|b\|_2 \sqrt{(\epsilon + \|E\|_2)^2 + \|E\|_2^2} \, . \tag{3.3.4}$$

Demostración. Veamos primero (3.3.3). La demostración está basada en el Corolario 3.2.4, que implica:

$$\begin{aligned}
\widetilde{x}_0 - x_0 &= \widetilde{A}^\dagger(b+h) - A^\dagger b \\
&= \left(\widetilde{A}^\dagger - A^\dagger\right)(b+h) + A^\dagger h \\
&= \left(A^\dagger \Theta_E + \Theta_F A^\dagger + \Theta_F A^\dagger \Theta_E\right)(b+h) + A^\dagger h \\
&= \left(A^\dagger \Theta_E + \Theta_F A^\dagger \Theta_E\right)(b+h) + \Theta_F x_0 + \Theta_F A^\dagger h + A^\dagger h \, .
\end{aligned}$$

Tomando norma 2 en ambos lados y teniendo en cuenta las propiedades de las normas, tenemos

$$\|\widetilde{x}_0 - x_0\|_2 \leq \|\Theta_F\|_2 \|x_0\|_2 + [\|\Theta_E\|_2 (1 + \|\Theta_F\|_2)(1+\epsilon) + \epsilon(1+\|\Theta_F\|_2)] \|A^\dagger\|_2 \|b\|_2 \, .$$

(3.3.3) se sigue del Lema 3.1.1 que implica $\|\Theta_E\|_2 \leq \alpha_E$ y $\|\Theta_F\|_2 \leq \alpha_F$.

Ahora, demostremos (3.3.4). Primero, observemos que

$$\begin{aligned}
\widetilde{r} - r &= h - \widetilde{A}\,\widetilde{x}_0 + Ax_0 \\
&= h - \widetilde{A}\widetilde{A}^\dagger \widetilde{b} + Ax_0 \\
&= (I - \widetilde{A}\widetilde{A}^\dagger) h + Ax_0 - \widetilde{A}\widetilde{A}^\dagger b \\
&= (I - \widetilde{A}\widetilde{A}^\dagger) h + Ax_0 - \widetilde{A}\widetilde{A}^\dagger (r + Ax_0) \\
&= (I - \widetilde{A}\widetilde{A}^\dagger)(h + Ax_0) - \widetilde{A}\widetilde{A}^\dagger r \, .
\end{aligned} \tag{3.3.5}$$

Observemos que los sumandos en (3.3.5) son vectores ortogonales, ya que $P_{\widetilde{A}} = \widetilde{A}\widetilde{A}^\dagger$, recordemos que $Ax_0 = P_A b$, $r = (I - P_A)b$ y el Teorema 3.2.3, para obtener (3.3.4) como a continuación

$$\begin{aligned}
\|\widetilde{r} - r\|_2^2 &= \|(I - P_{\widetilde{A}})(h + P_A b)\|_2^2 + \|P_{\widetilde{A}}(I - P_A)b\|_2^2 \\
&\leq \left(\|h\|_2 + \|(I - P_{\widetilde{A}})P_A b\|_2\right)^2 + \|P_{\widetilde{A}}(I - P_A)b\|_2^2 \\
&\leq (\epsilon\|b\|_2 + \|E\|_2\|b\|_2)^2 + \|E\|_2^2 \|b\|_2^2 \, .
\end{aligned}$$

□

Si A tiene rango completo por filas o columnas la cota (3.3.3) se simplifica en el sentido explicado en las partes (d) y (e) de la Observación 3.2.6. Si A tiene rango completo por filas, entonces $\widetilde{r} = r = 0$ y $\|\widetilde{r} - r\|_2 = 0$.

Las cotas en el Teorema 3.3.1 mejoran significativamente las cotas clásicas para el cambio relativo de las soluciones de mínima longitud y para los residuos del problema de mínimos cuadrados bajo perturbaciones aditivas $\widetilde{A} = A + \Delta A$ [74, Teorema 5.1]. Con el objetivo de comparar, enunciamos estas cotas clásicas de perturbaciones como aparecen en [5, Teorema 1.4.6] para el caso $rango\,(A) = rango\,(\widetilde{A})$ y $\eta := \kappa_2(A)\,\|\Delta A\|_2/\|A\|_2 < 1$. Sea $\widetilde{x}_0$ la solución mínima del problema de mínimos cuadrados $\min_{x\in\mathbb{R}^n} \|(b + \Delta b) - (A + \Delta A)x\|_2$ y x_0 la solución mínima de $\min_{x\in\mathbb{R}^n} \|b - Ax\|_2$, y sean $\widetilde{r} := (b + \Delta b) - (A + \Delta A)\widetilde{x}_0$ y $r := b - Ax_0$. Entonces, si suponemos que $x_0 \neq 0$ y definimos $\epsilon_A := \|\Delta A\|_2/\|A\|_2$ y $\epsilon_b := \|\Delta b\|_2/\|b\|_2$, por el Teorema 1.4.6 de [5] tenemos

$$\frac{\|\widetilde{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \frac{1}{1-\eta}\left(2\,\kappa_2(A)\,\epsilon_A + \frac{\|A^\dagger\|_2\,\|b\|_2}{\|x_0\|_2}\,\epsilon_b + \kappa_2(A)^2\,\frac{\|r\|_2}{\|A\|_2\,\|x_0\|_2}\,\epsilon_A\right) \text{ y} \tag{3.3.6}$$

$$\frac{\|\widetilde{r} - r\|_2}{\|b\|_2} \leq \left(\frac{\|A\|_2\|x_0\|_2}{\|b\|_2}\,\epsilon_A + \epsilon_b + \kappa_2(A)\,\frac{\|r\|_2}{\|b\|_2}\,\epsilon_A\right). \tag{3.3.7}$$

En la ecuación (3.3.7), es conveniente recordar que $(\|A\|_2\|x_0\|_2)/\|b\|_2 \leq \kappa_2(A)$ y $\|r\|_2 \leq \|b\|_2$. Luego, observemos que la cota para $\|\widetilde{r} - r\|_2/\|b\|_2$ en (3.3.7) incluye términos que pueden ser muy grandes aunque ϵ_A y ϵ_b sean muy pequeños. Esto sucede si $\kappa_2(A)$ es grande y si $\|r\|_2 \neq 0$ no es muy pequeño. En cambio, si $\|E\|_2$ y ϵ son muy pequeños, entonces la cota para $\|\widetilde{r} - r\|_2/\|b\|_2$ que se obtiene de (3.3.4) siempre es muy pequeña. Con respecto a las cotas para $\|\widetilde{x}_0 - x_0\|_2/\|x_0\|_2$: la cota en (3.3.6) aumenta las perturbaciones en los datos al menos por un factor $\kappa_2(A)$ y el aumento puede ser más grande bajo ciertas condiciones. Adicionalmente, (3.3.6) incluye el factor $\|A^\dagger\|_2\,\|b\|_2/\|x_0\|_2$, el cual es el único factor potencialmente grande en la cota que se obtiene de (3.3.3). Mostraremos en la Sección 3.3.1 que $\|A^\dagger\|_2\,\|b\|_2/\|x_0\|_2$ es un número moderado excepto para elecciones muy particulares de b. Por lo tanto, (3.3.3) siempre mejora (3.3.6) y, si $\|E\|_2$, $\|F\|_2$, y ϵ son muy pequeños, entonces (3.3.3) produce cotas muy pequeñas para $\|\widetilde{x}_0 - x_0\|_2/\|x_0\|_2$ para casi todo b.

Las cotas en el Teorema 3.3.1 no pueden ser aplicadas directamente debido a la presencia de $\widehat{E}$ y $\widehat{F}$. El Corolario 3.3.2 resuelve esta deficiencia restringiendo la magnitud de las perturbaciones y usando la ecuación (3.2.10). El Corolario 3.3.2 se

demuestra directamente del Teorema 3.3.1 y es enunciado de manera conveniente para su uso en la Sección 3.4.

Corolario 3.3.2 *Con la misma notación e hipótesis del Teorema 3.3.1, y suponiendo que* $\|E\|_2 \leq \mu < 1$, $\|F\|_2 \leq \nu < 1$, $x_0 \neq 0$ *y* $b \neq 0$. *Definamos*

$$\theta_\mu := \mu \sqrt{1 + \frac{1}{(1-\mu)^2}} \quad y \quad \theta_\nu := \nu \sqrt{1 + \frac{1}{(1-\nu)^2}} \,. \tag{3.3.8}$$

Entonces las siguientes cotas se cumplen

$$\frac{\|\widetilde{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \theta_\nu + [\theta_\mu(1+\theta_\nu)(1+\epsilon) + \epsilon(1+\theta_\nu)] \, \frac{\|A^\dagger\|_2 \|b\|_2}{\|x_0\|_2} \,, \tag{3.3.9}$$

$$\frac{\|\widetilde{r} - r\|_2}{\|b\|_2} \leq \sqrt{(\epsilon+\mu)^2 + \mu^2} \,. \tag{3.3.10}$$

La cota (3.3.9) *se cumple a primer orden en* ϵ, μ *y* ν

$$\frac{\|\widetilde{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \sqrt{2}\,\nu + \left(\epsilon + \sqrt{2}\,\mu\right) \frac{\|A^\dagger\|_2 \|b\|_2}{\|x_0\|_2} + \; h.o.t \,, \tag{3.3.11}$$

donde h.o.t significa términos de orden superior (high order terms) en ϵ, μ *y* ν.

Probaremos en la Sección 3.3.2 que, a primer orden, la cota de perturbaciones para $\|\widetilde{x}_0 - x_0\|_2 / \|x_0\|_2$ que se obtiene del Teorema 3.3.1 es óptima. Es decir, puede ser alcanzada salvo una constante. En este contexto, es interesante observar que otro procedimiento para obtener cotas de perturbaciones multiplicativas para el problema de mínimos cuadrados es utilizar el Teorema 3.2.5, como describimos a continuación: sabemos que $\widetilde{x}_0 - x_0 = \left(\widetilde{A}^\dagger - A^\dagger\right)(b+h) + A^\dagger h$, así que $\|\widetilde{x}_0 - x_0\|_2 \leq \|\widetilde{A}^\dagger - A^\dagger\|_2 (\|b\|_2 + \|h\|_2) + \|A^\dagger\|_2 \|h\|_2$, esto puede ser combinado con el Teorema 3.2.5-(b) y $\|h\|_2 \leq \epsilon \|b\|_2$ para obtener una cota para $\|\widetilde{x}_0 - x_0\|_2 / \|x_0\|_2$. La cota obtenida (la llamaremos Γ_{LP}) incluye en todos sus términos el factor $\|A^\dagger\|_2 \, \|b\|_2 / \|x_0\|_2$. Esto contrasta con la cota obtenida de (3.3.3) (lo llamaremos Γ_{LS}) la cual tiene un término que es simplemente α_F (ver también los términos θ_ν en (3.3.9) ó $\sqrt{2}\,\nu$ en (3.3.11)). Así, Γ_{LP} es mucho mayor que Γ_{LS} si $\max\{\|E\|_2, \epsilon\} \ll \|F\|_2$ y $\|A^\dagger\|_2 \|b\|_2 / \|x_0\|_2$ es grande (esto se puede comprobar fácilmente a primer orden). Adicionalmente, podemos demostrar a primer orden que siempre se cumple que $\Gamma_{LS} \leq 2\,\Gamma_{LP}$ y si $\|A^\dagger\|_2 \|b\|_2 / \|x_0\|_2 \geq 2$ entonces $\Gamma_{LS} \leq \Gamma_{LP}$. Entonces, podemos concluir que el método considerado en el Teorema 3.3.1 es mejor que el usado en el Teorema 3.2.5.

Finalmente, observemos que todos los resultados en esta sección, igual que en la Sección 3.2, son válidos para cualquier valor de m y n, esto es, si $m \geq n$ ó $m < n$. Así que, también son válidos para perturbaciones multiplicativas de soluciones de sistemas lineales infradeterminados.

3.3.1. ¿Por qué el factor $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$ es pequeño?

Esta sección está relacionada con [32, Sección 3.2] (ver también Sección 2.3), que considera el mismo problema para una matriz no singular A. Aunque el hecho que $A \in \mathbb{R}^{m\times n}$ sea rectangular, obliga a realizar modificaciones no triviales, las conclusiones más importantes permanecen igual. El Teorema 3.3.1 y el Corolario 3.3.2 prueban que la sensibilidad de la solución de mínima longitud, $x_0 = A^\dagger b$ del problema de mínimos cuadrados bajo perturbaciones multiplicativas está regida por $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$. Esta cantidad es justamente el número de condición del problema de mínimos cuadrados, cuando sólamente el lado derecho b de $\min_{x\in\mathbb{R}^n} \|Ax - b\|_2$ es perturbado, como lo demostramos en la ecuación (2.3.9). Más precisamente, es fácil demostrar que si $x_0 = A^\dagger b$, entonces

$$\frac{\|A^\dagger\|_2 \, \|b\|_2}{\|x_0\|_2} = \lim_{\epsilon\to 0} \sup \left\{ \frac{\|\widetilde{x}_0 - x_0\|_2}{\epsilon \|x_0\|_2} : \widetilde{x}_0 = A^\dagger \, (b+h), \;\; \|h\|_2 \le \epsilon \, \|b\|_2 \right\} .$$

Por lo tanto, el Teorema 3.3.1 prueba básicamente que las perturbaciones multiplicativas tienen un efecto sobre la solución de mínima longitud del problema de mínimos cuadrados similar a perturbar sólamente el lado derecho b.

Notemos que $1 \le \|A^\dagger\|_2 \|b\|_2/\|x_0\|_2$, pero, en general, $\|A^\dagger\|_2 \|b\|_2/\|x_0\|_2 \not\le \kappa_2(A)$, contrastando con el caso cuando A es no singular[3] [32, Sección 3.2]. Sin embargo, (3.3.6) muestra que

$$\kappa_{LS}(A,b) := \left(2\,\kappa_2(A) + \frac{\|A^\dagger\|_2 \, \|b\|_2}{\|x_0\|_2} + \kappa_2^2(A) \, \frac{\|r\|_2}{\|A\|_2 \, \|x_0\|_2} \right)$$

puede ser considerado como un número de condición para el problema de mínimos cuadrados bajo perturbaciones aditivas muy pequeñas de A y b (en efecto, se ha demostrado en [74, Sección 6] que la cota (3.3.6) es aproximadamente alcanzada a primer orden en las perturbaciones), y $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2 \le \kappa_{LS}(A,b)$. Pero el primer punto clave en esta sección es mostrar que *si A es fija*, entonces $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$ es un número moderado *para la mayoría de los vectores* b, incluso si $\kappa_2(A) \gg 1$ y así $\kappa_{LS}(A,b) \gg 1$, lo cual implica que $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2 \ll \kappa_{LS}(A,b)$ para la mayoría de los problemas de mínimos cuadrados mal condicionados cuya matriz de coeficientes es A. Sin embargo, esto no es suficiente para nuestro propósito, ya que si $rango(A) < m$, entonces para la mayoría de vectores b el ángulo agudo $\theta(b, \mathcal{R}(A))$ entre b y el espacio columna de A no es pequeño, lo cual es equivalente a decir que el residuo

[3] Si A es no singular o, si $Ax_0 = b$, entonces $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2 \le \kappa_2(A)$. Sin embargo, consideremos $A = [1 \;\; 0; 0 \;\; 1; 0 \;\; 0] \in \mathbb{R}^{3\times 2}$ y $b = [\eta; 0; 1] \in \mathbb{R}^{3\times 1}$. En este caso $x_0 = [\eta; 0] \in \mathbb{R}^{2\times 1}$ y $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2 = \sqrt{|\eta|^2 + 1}/|\eta|$ tiende a ∞ si $\eta \to 0$ mientras $\kappa_2(A) = 1$.

relativo $\|Ax_0-b\|_2/\|b\|_2 = \sin\theta(b,\mathcal{R}(A))$ no es pequeño. Pero muy frecuentemente en la práctica, los problemas de mínimos cuadrados tienen residuos relativos pequeños, dado que los problemas corresponden a sistemas lineales inconsistentes $Ax \approx b$ que están cercanos de ser consistentes. Por lo tanto, el segundo punto importante en esta sección es que, *si fijamos* A para considerar todos los vectores b tales que $\Upsilon = \theta(b,\mathcal{R}(A)) < \pi/2$ *también fijo*, podemos demostrar que para la mayoría *de estos vectores* b el factor $\|A^\dagger\|_2\,\|b\|_2/\|x_0\|_2$ es un número moderado mucho más pequeño que $\kappa_{LS}(A,b)$ cuando A está muy mal condicionada.

Para explicar las propiedades mencionadas anteriormente, supongamos que $rango\,(A) = r$ y sea $A = U\Sigma V^T$ la SVD de A, donde $U \in \mathbb{R}^{m\times r}$ y $V \in \mathbb{R}^{n\times r}$ tienen columnas ortonormales, $\Sigma = \mathrm{diag}(\sigma_1,\ldots,\sigma_r) \in \mathbb{R}^{r\times r}$, y $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Observemos que $\|x_0\|_2 = \|A^\dagger b\|_2 = \|\Sigma^{-1}U^T b\|_2 \geq |u_r^T b|/\sigma_r$ y

$$\frac{\|A^\dagger\|_2\,\|b\|_2}{\|x_0\|_2} = \frac{\|b\|_2}{\sigma_r\|x_0\|_2} \leq \frac{\|b\|_2}{|u_r^T b|} = \frac{1}{\cos\,\theta(u_r,b)}, \tag{3.3.12}$$

donde u_r es la última columna de U y $\theta(u_r,b)$ es el ángulo agudo entre u_r y b. Note que la cota sobre $\|A^\dagger\|_2\,\|b\|_2/\|x_0\|_2$ en (3.3.12) puede ser grande sólo si b es "casi" ortogonal a u_r. Por ejemplo, si A es una matriz fija extremadamente mal condicionada (por ejemplo $\kappa_2(A) = 10^{1000}$) y b es considerado como un vector aleatorio cuya dirección está uniformemente distribuida en todo el espacio, entonces la probabilidad que $0 \leq \theta(u_r,b) \leq \pi/2 - 10^{-6}$ es aproximadamente $1 - 10^{-6}$. Observemos que la condición $0 \leq \theta(u_r,b) \leq \pi/2 - 10^{-6}$ implica que $\|A^\dagger\|_2\,\|b\|_2/\|x_0\|_2 \lesssim 10^6$, lo cual es un número moderado comparado a 10^{1000}. En particular, si los parámetros perturbativos μ, ν y ϵ en el Corolario 3.3.2 son 10^{-16}, entonces $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2 \lesssim 10^6$ proporciona una excelente cota para la variación de la solución de mínima longitud del problema de mínimos cuadrados. Incluso, es posible que $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ sea moderado aunque $\cos\,\theta(u_r,b) \approx 0$. Esto puede verse como una extensión del caso de matrices no singulares a matrices generales del resultado original presentado por Chan y Foulser (Teorema 2.3.2).

En el argumento anterior, el vector b puede estar en cualquier lugar en el espacio. Luego, consideremos vectores b tales que $\Upsilon = \theta(b,\mathcal{R}(A)) < \pi/2$ se mantiene constante. Describiraremos todos estos vectores a continuación: sea $y \in \mathbb{R}^r$ cualquier vector y sea $U_\perp \in \mathbb{R}^{m\times(m-r)}$ tal que $[\,U\;U_\perp\,] \in \mathbb{R}^{m\times m}$ es unitaria. Entonces elegimos cualquier $z \in \mathbb{R}^{m-r}$ tal que $\|z\|_2 = \|y\|_2\,\tan\Upsilon$, y definimos $b = Uy + U_\perp z$. Sabemos que $\Upsilon = \theta(b,\mathcal{R}(A))$, ya que $\mathcal{R}(U) = \mathcal{R}(A)$. Además, de (3.3.12), es fácil demostrar

que estos vectores b satisfacen

$$\frac{\|A^\dagger\|_2\,\|b\|_2}{\|x_0\|_2} = \frac{\|b\|_2}{\sigma_r\|x_0\|_2} \leq \frac{\|b\|_2}{|u_r^T b|} = \frac{\sqrt{1+\tan^2\Upsilon}}{\cos\,\theta(e_r,y)} = \frac{1}{(\cos\,\Upsilon)\cdot(\cos\,\theta(e_r,y)\,)}, \quad (3.3.13)$$

donde e_r es la r-ésima columna de I_r. La cota en (3.3.13) es una cantidad "geométrica" que no depende de $\kappa_2(A)$, si suponemos que Υ no es muy cercano a $\pi/2$, y que es un número moderado para la mayoría de los vectores y, es decir, para la mayoría de los vectores b tales que $\Upsilon = \theta(b, \mathcal{R}(A))$.

Finalmente, discutimos una interesante relación entre $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ con el término de $\kappa_{LS}(A,b)$ que depende de $\kappa_2^2(A)$. Note que este término puede acotarse como sigue

$$\Phi := \kappa_2^2(A)\,\frac{\|r\|_2}{\|A\|_2\,\|x_0\|_2} \;=\; \kappa_2(A)\,\frac{\|A^\dagger\|_2\,\|r\|_2}{\|x_0\|_2} \;\leq\; \kappa_2(A)\,\frac{\|A^\dagger\|_2\,\|b\|_2}{\|x_0\|_2}\,. \quad (3.3.14)$$

En relación a nuestra discusión $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ es un número moderado para la mayoría de los vectores b. Por lo tanto, Φ es acotado superiormente por un número moderado de veces de $\kappa_2(A)$ para la mayoría de los vectores b y como consecuencia, $\kappa_2^2(A)$ sólamente afecta la sensibilidad del problema de mínimos cuadrados en situaciones muy particulares. Además, Φ puede escribirse como:

$$\left(\kappa_2(A)\,\frac{\|A^\dagger\|_2\,\|b\|_2}{\|x_0\|_2}\right)\,\frac{\|r\|_2}{\|b\|_2} \;= \Phi\,, \quad (3.3.15)$$

lo cual implica que para residuos relativos suficientemente grandes (pensar, por ejemplo, en $\|r\|_2/\|b\|_2 \geq 10^{-3}$) y matrices A muy mal condicionadas, tenemos $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2 \ll \Phi \leq \kappa_{LS}(A,b)$, incluso si $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ es grande.

3.3.2. El número de condición para perturbaciones multiplicativas del problema de mínimos cuadrados

En esta sección demostramos que $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ es el número de condición bajo perturbaciones multiplicativas del problema de mínimos cuadrados salvo una constante. El lector debe notar que, por simplicidad, consideramos en nuestra definición de número de condición que la variación relativa de b y las perturbaciones multiplicativas izquierdas y derechas tienen el mismo orden.

Teorema 3.3.3 *Usando la misma notación e hipótesis del Corolario 3.3.2 con los parámetros μ, ν y ϵ iguales a η, definamos el número de condición*

$$\kappa_{LS}^{(M)}(A,b) := \lim_{\eta\to 0}\sup\left\{ \frac{\|\widetilde{x}_0 - x_0\|_2}{\eta\,\|x_0\|_2} \;:\; \widetilde{x}_0 = [(I+E)A(I+F)]^{\dagger}\,(b+h), \right. \\ \left. \|E\|_2 \le \eta,\ \|F\|_2 \le \eta,\ \|h\|_2 \le \eta\|b\|_2 \right\}.$$

Entonces,

$$\frac{1}{1+2\sqrt{2}}\,\kappa_{LS}^{(M)}(A,b) \le \frac{\|A^{\dagger}\|_2\,\|b\|_2}{\|x_0\|_2} \le \kappa_{LS}^{(M)}(A,b)\,. \tag{3.3.16}$$

Demostración. De la ecuación (3.3.11) y de $1 \le \|A^{\dagger}\|_2\,\|b\|_2/\|x_0\|_2$, tenemos

$$\frac{\|\widetilde{x}_0 - x_0\|_2}{\eta\,\|x_0\|_2} \le \sqrt{2} + \left(1+\sqrt{2}\right)\frac{\|A^{\dagger}\|_2\|b\|_2}{\|x_0\|_2} + O(\eta) \le \left(1+2\sqrt{2}\right)\frac{\|A^{\dagger}\|_2\|b\|_2}{\|x_0\|_2} + O(\eta)\,,$$

lo cual implica la desigualdad izquierda de (3.3.16). Para demostrar la desigualdad del lado derecho elegimos una perturbación tal que $E = 0$, $F = 0$ y $h = \eta w$, donde $\|w\|_2 = \|b\|_2$ y $\|A^{\dagger}w\|_2 = \|A^{\dagger}\|_2\|w\|_2$. Para esta perturbación $\|\widetilde{x}_0 - x_0\|_2/\|x_0\|_2 = \|A^{\dagger}h\|_2/\|x_0\|_2 = \eta\,\|A^{\dagger}\|_2\|b\|_2/\|x_0\|_2$. Por lo tanto, la expresión "sup" que aparece en la definición de $\kappa_{LS}^{(M)}(A,b)$ implica que $\|A^{\dagger}\|_2\|b\|_2/\|x_0\|_2 \le \kappa_{LS}^{(M)}(A,b)$. □

3.3.3. Cotas de perturbaciones multiplicativas para otras soluciones del problema de mínimos cuadrados

Cotas para el cambio en las soluciones de problemas de mínimos cuadrados se pueden obtener fácilmente a partir del Teorema 3.3.1 y Teorema 3.2.3-(c) que son una pequeña modificación de la ecuación (3.3.3). Dado que el residuo de un problema de mínimos cuadrados es el mismo para todas sus soluciones, no es necesario considerar nuevas cotas perturbativas para el residuo.

Teorema 3.3.4 *Si $y \in \mathbb{R}^n$ es una solución del problema de mínimos cuadrados* (3.3.1)*, entonces existe una solución $\widetilde{y} \in \mathbb{R}^n$ del problema de mínimos cuadrados* (3.3.2) *tal que*

$$\|\widetilde{y} - y\|_2 \le\ (\alpha_F + \|F\|_2)\,\|y\|_2 + [\alpha_E\,(1+\alpha_F)\,(1+\epsilon) + \epsilon\,(1+\alpha_F)\,]\ \|A^{\dagger}\|_2\,\|b\|_2\,,$$

donde α_E, α_F y ϵ se definen como en el enunciado del Teorema 3.3.1.

Demostración. Dado y, existe un vector $z \in \mathbb{R}^n$ tal que $y = x_0 + P_{\mathcal{N}(A)}z$, donde x_0 es la solución de mínima longitud de (3.3.1). Recordemos también que $\|y\|_2^2 = \|x_0\|_2^2 + \|P_{\mathcal{N}(A)}z\|_2^2$ y por lo tanto, $\|P_{\mathcal{N}(A)}z\|_2 \leq \|y\|_2$. Elegimos la siguiente solución de (3.3.2), $\widetilde{y} = \widetilde{x}_0 + P_{\mathcal{N}(\widetilde{A})}P_{\mathcal{N}(A)}z$, donde $\widetilde{x}_0$ es la solución mínima de (3.3.2). Por lo tanto

$$\|\widetilde{y} - y\|_2 \leq \|\widetilde{x}_0 - x_0\|_2 + \|(P_{\mathcal{N}(\widetilde{A})} - P_{\mathcal{N}(A)})P_{\mathcal{N}(A)}z\|_2 \leq \|\widetilde{x}_0 - x_0\|_2 + \|F\|_2\|y\|_2\,,$$

donde hemos utilizado el Teorema 3.2.3-(c). Ahora, usamos (3.3.3) y $\|x_0\|_2 \leq \|y\|_2$ para obtener el resultado. □

Observemos que el cambio relativo $\|\widetilde{y} - y\|_2/\|y\|_2$ está acotado por $\text{máx}\{1, \|A^\dagger\|_2\|b\|_2/\|y\|_2\}$, que es menor o igual que $\|A^\dagger\|_2\,\|b\|_2/\|x_0\|_2$. Por lo tanto, la solución de mínima longitud es la más sensible de las soluciones bajo perturbaciones multiplicativas.

3.4. Perturbación del problema de mínimos cuadrados en los factores

Como explicamos en la Introducción, presentamos en la Sección 3.5 un algoritmo preciso para la solución del problema de mínimos cuadrados $\text{mín}_{x\in\mathbb{R}^n}\,\|b - Ax\|_2$ utilizando una RRD precisa XDY de A, el Algoritmo 3.5.1. El análisis de errores del Algoritmo 3.5.1 lo presentamos en el Teorema 3.5.2 y muestra que la solución calculada es la solución exacta del problema de mínimos cuadrados correspondiente a una RRD con factores cercanos a $(X + \Delta X)(D + \Delta D)(Y + \Delta Y)$, donde las perturbaciones son "normwise" para los factores bien condicionados X e Y, y "componentwise" para los elementos diagonales y potencialmente mal condicionados del factor D. Por lo tanto, es necesario desarrollar cotas perturbativas para la solución del problema de mínimos cuadrados cuya matriz de coeficientes está expresada como una RRD bajo perturbaciones en los factores. Esto es presentado en el Teorema 3.4.1, y su demostración se basa en escribir las perturbaciones en los factores como una perturbación multiplicativa de toda la matriz.

Recordemos que, por el Lema 3.1.2-(c), si $A = XDY$ es una RRD de A, entonces $A^\dagger = Y^\dagger D^{-1}X^\dagger$. Por lo tanto, la solución de mínima longitud del problema de mínimos cuadrados $\text{mín}_{x\in\mathbb{R}^n}\,\|b - XDYx\|_2$ es $x_0 = Y^\dagger D^{-1}X^\dagger b$.

Teorema 3.4.1 *Sean $X \in \mathbb{R}^{m\times r}$, $D \in \mathbb{R}^{r\times r}$ e $Y \in \mathbb{R}^{r\times n}$ tales que rango(X) = rango(Y) = r, D es diagonal y no singular y $b \in \mathbb{R}^m$. Sea x_0 la solución de*

mínima longitud de $\min_{x\in\mathbb{R}^n} \|b - XDY\,x\|_2$ *y* $\widetilde{x}_0$ *la solución de mínima longitud de* $\min_{x\in\mathbb{R}^n} \|(b+h) - (X+\delta X)(D+\delta D)(Y+\delta Y)\,x\|_2$*, donde* $\|\delta X\|_2 \le \alpha\|X\|_2$*,* $\|\delta Y\|_2 \le \beta\|Y\|_2$*,* $|\delta D| \le \rho|D|$*, y* $\|h\|_2 \le \epsilon\|b\|_2$*. Sean* $r = b - XDY\,x_0$ *y* $\widetilde{r} = (b+h) - (X+\delta X)(D+\delta D)(Y+\delta Y)\,\widetilde{x}_0$*. Supongamos que*

$$\mu := \alpha\,\kappa_2(X) < 1 \quad y \quad \nu := [\beta + \rho(1+\beta)]\kappa_2(Y) < 1, \tag{3.4.1}$$

y definamos los parámetros θ_μ *y* θ_ν *como en la ecuación* (3.3.8)*. Entonces, la cota* (3.3.9) *se satisface con* $A^\dagger$ *reemplazada por* $Y^\dagger D^{-1} X^\dagger$*, la cota* (3.3.10) *se satisface y a primer orden en* α, β, ρ *y* ϵ *tenemos*

$$\frac{\|\widetilde{x}_0 - x_0\|_2}{\|x_0\|_2} \le \sqrt{2}\,(\beta+\rho)\,\kappa_2(Y) + \Big(\epsilon + \sqrt{2}\,\alpha\,\kappa_2(X)\Big)\,\frac{\|Y^\dagger D^{-1} X^\dagger\|_2 \|b\|_2}{\|x_0\|_2} + h.o.t. \tag{3.4.2}$$

Demostración. Sean $A = XDY$ y $\widetilde{A} = (X+\delta X)(D+\delta D)(Y+\delta Y)$. Escribamos $\widetilde{A}$ como una perturbación multiplicativa de A:

$$\begin{aligned}
\widetilde{A} &= (I + \delta X X^\dagger)\,XD\,(I + D^{-1}\,\delta D)Y(I + Y^\dagger \delta Y) \\
&= (I + \delta X X^\dagger)\,XDY\,(I + Y^\dagger\,D^{-1}\,\delta D\,Y)\,(I + Y^\dagger \delta Y) \\
&=: (I+E)A(I+F),
\end{aligned}$$

donde $E = \delta X X^\dagger$ y $F = Y^\dagger \delta Y + Y^\dagger D^{-1}\,\delta D\,Y + Y^\dagger D^{-1}\,\delta D\,\delta Y$. Luego, tomando en cuenta que $\|\delta D\,D^{-1}\|_2 \le \rho$, tenemos

$$\|E\|_2 \le \alpha\,\kappa_2(X) = \mu < 1, \quad \|F\|_2 \le [\beta + \rho(1+\beta)]\,\kappa_2(Y) = \nu < 1,$$

y el Teorema 3.4.1 se obtiene inmediatamente del Corolario 3.3.2. □

Dado que los factores X e Y de una RRD están bien condicionados, de la ecuación (3.4.2) tenemos que la sensitividad respecto a perturbaciones de los factores de la solución de mínima longitud del problema de mínimos cuadrados $\min_{x\in\mathbb{R}^n} \|b - XDY\,x\|_2$ está nuevamente controlada por $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$, donde $A = XDY$, que es un número moderado para la mayoría de los vectores b (ver Sección 3.3.1). Note que el Teorema 3.4.1 es válido si $m \ge n$ o si $m < n$, es decir, para problemas de mínimos cuadrados o para sistemas de ecuaciones lineales infradeterminados, ya que el Corolario 3.3.2 se cumple en ambos casos.

3.5. Algoritmo y análisis de errores

En esta sección presentamos el Algoritmo 3.5.1 para resolver el problema de mínimos cuadrados $\min_{x\in\mathbb{R}^n} \|b - A\,x\|_2$ y demostramos que calcula la solución de

mínima longitud con error relativo acotado por $O(\mathtt{u})\,\|A^\dagger\|_2\,\|b\|_2/\|x_0\|_2$, lo cual es simplemente $O(\mathtt{u})$ para la mayoría de vectores b de acuerdo con lo discutido en la Sección 3.3.1. El primer paso del algoritmo calcula una RRD precisa de $A = XDY \in \mathbb{R}^{m\times n}$ en el sentido de la Definición 2.4.2, lo cual es posible para muchas clases de matrices estructuradas como hemos dicho en la Introducción. Los siguientes pasos del Algoritmo 3.5.1 están basados en el hecho de que, acorde al Lema 3.1.2-(c), la solución de mínima longitud es $x_0 = Y^\dagger(D^{-1}(X^\dagger b))$ y en las siguientes observaciones: (1) $s = X^\dagger b$ es la solución única del problema de mínimos cuadrados *de rango completo por columnas* $\text{mín}_{x\in\mathbb{R}^r}\,\|b - X\,x\|_2$; (2) $w = D^{-1}(X^\dagger b)$ es la solución única del sistema lineal $Dw = s$; y (3) $Y^\dagger(D^{-1}(X^\dagger b))$ es la solución única del sistema lineal indeterminado con *rango completo por filas* $Yx = w$. Observemos que este método es válido si $m \geq n$ y si $m < n$. Por lo tanto, en el último caso y, si $rango\,(A) = m$, el método resuelve de manera precisa el sistema lineal infradeterminado $Ax = b$.

La solución mínima x_0 del sistema infraterminado $Yx = x_2$ es calculado vía el Q-Método descrito en [43, Sección 21.1] y que recordamos brevemente a continuación. La idea es calcular a primer orden una factorización QR *reducida* de $Y^T = W R_Y \in \mathbb{R}^{n\times r}$ utilizando el algoritmo de Householder, donde $W \in \mathbb{R}^{n\times r}$ satisface $W^TW = I_r$ y $R_Y \in \mathbb{R}^{r\times r}$ es triangular superior y no singular. Así, $Y = R_Y^T W^T \in \mathbb{R}^{r\times n}$ y el Lema 3.1.2-(c) implican que $Y^\dagger = (W^T)^\dagger\,(R_Y^T)^\dagger = W\,R_Y^{-T}$, donde R_Y^{-T} denota la inversa de R_Y^T. Finalmente, $x_0 = W\,(R_Y^{-T}x_2)$ y $R_Y^{-T}x_2$ es calculado resolviendo el sistema triangular $R_Y^T x = x_2$ por sustitución progresiva. En la práctica, es importante notar que el factor W no se requiere explícitamente, sólamente necesitamos la habilidad de multiplicar W veces un vector y esto puede hacerse multiplicando las permutaciones de Householder resultantes de la factorización QR de Y^T por $[(R_Y^{-T}x_2)^T, 0]^T \in \mathbb{R}^n$.

Estamos ahora en posición de enunciar el Algoritmo 3.5.1.

Algoritmo 3.5.1 (Solución precisa del problema de mínimos cuadrados vía la RRD)

`Input:` $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$

`Output:` x_0 la solución de $\text{mín}_{x\in\mathbb{R}^n}\,\|b - A\,x\|_2$

`Paso 1:` Calcular una RRD precisa de $A = XDY$ en el sentido de la Definición 2.4.2, donde $X \in \mathbb{R}^{m\times r}$, $D \in \mathbb{R}^{r\times r}$ es diagonal, $Y \in \mathbb{R}^{r\times n}$ y $rango\,(A) = rango\,(X) = rango\,(Y) = rango\,(D) = r$.

`Paso 2:` Calcular la solución única s de $\min_{x\in\mathbb{R}^r} \|b - X\,x\|_2$ usando la factorización QR de Householder de X.

`Paso 3:` Calcular la solución única w del sistema lineal diagonal $D\,w = s$ como $w_i = s_i/d_{ii}$, $i = 1, \ldots, r$.

`Paso 4:` Calcular la solución de mínima longitud x_0 de $Yx = w$ usando el Q-método, es decir, vía la factorización QR de Householder de Y^T.

Antes de realizar el análisis de errores del Algoritmo 3.5.1, observemos que por el teorema de perturbaciones de Weyl [69], las diferencias entre los valores singulares ordenados de X y $\widehat{X}$ están acotadas como $|\sigma_i(\widehat{X}) - \sigma_i(X)| \leq \|\widehat{X} - X\|_2 \leq \mathtt{u}\,p(m,n)\,\|X\|_2$, para $i = 1 : r$. Por lo tanto, $|\sigma_i(\widehat{X}) - \sigma_i(X)|/\sigma_i(X) \leq \mathtt{u}\,p(m,n)\,\kappa_2(X)$, para $i = 1 : r$. Análogamente podemos acotar las diferencias entre los valores singulares ordenados de Y e $\widehat{Y}$. Como consecuencia, la condición (2.4.3) implica que $rango\,(X) = rango\,(\widehat{X}) = r$, $rango\,(Y) = rango\,(\widehat{Y}) = r$ y

$$\frac{\kappa_2(X)}{3} \leq \kappa_2(\widehat{X}) \leq 3\,\kappa_2(X) \quad \text{y} \quad \frac{\kappa_2(Y)}{3} \leq \kappa_2(\widehat{Y}) \leq 3\,\kappa_2(Y)\,. \tag{3.5.1}$$

La ecuación (3.5.1) nos permite usar indistintamente $\kappa_2(X)$, $\kappa_2(Y)$ ó $\kappa_2(\widehat{X})$, $\kappa_2(\widehat{Y})$ en las cotas de error obtenidas con el coste de modificar ligeramente algunas de las constantes involucradas.

El coste computacional en el `Paso 1` del Algoritmo 3.5.1 depende del tipo específico de matrices y del algoritmo usado entre los mencionados en la introducción. De cualquier manera, todos estos algoritmos tienen un coste de $O(mn^2)$ flops si $m \geq n$ y $O(m^2n)$ flops si $m < n$. El término principal del coste de los `Pasos 2, 3` y `4` es $2r^2(m - r/3)$, r y $2r^2(n - r/3)$ flops, respectivamente. Dado que $r \leq \min\{m, n\}$, el coste total del Algoritmo 3.5.1 es $O(mn^2)$ flops si $m \geq n$ y $O(m^2n)$ flops si $m < n$.

Los errores de redondeo regresivos cometidos por el Algoritmo 3.5.1 son analizados en el Teorema 3.5.2. Usaremos la siguiente notación introducida en [43, Secciones 3.1 y 3.4]

$$\gamma_n := \frac{n\mathtt{u}}{1 - n\mathtt{u}} \quad \text{and} \quad \widetilde{\gamma}_n := \frac{cn\mathtt{u}}{1 - cn\mathtt{u}}, \tag{3.5.2}$$

donde c denota una pequeña constante entera cuyo valor exacto no es esencial en el análisis. Antes de enunciar y demostrar el Teorema 3.5.2, comentamos la necesidad

y el significado de las hipótesis usadas en el teorema. Primero, asumimos que los factores $\widehat{X}$, $\widehat{D}$ e $\widehat{Y}$ calculados en el `Paso 1` en aritmética en coma flotante satisfacen (2.4.1), (2.4.2) y (2.4.3) , lo cual implica que $rango\,(X) = rango\,(\widehat{X}) = r$, $rango\,(D) = rango\,(\widehat{D}) = r$, $rango\,(Y) = rango\,(\widehat{Y}) = r$ y (3.5.1). Por lo tanto, podemos usar $\kappa_2(X)$ y $\kappa_2(Y)$ en los errores de los `Pasos 2` y `4` en vez de $\kappa_2(\widehat{X})$ y $\kappa_2(\widehat{Y})$ al coste de no prestar atención a los valores exactos de las constantes numéricas en las cotas de errores. La hipótesis $\text{máx}\{\kappa_2(X), \kappa_2(Y)\}\,\sqrt{r}\,\widetilde{\gamma}_{r\,\text{máx}\{m,n\}} < 1$ garantiza que los errores regresivos $\Delta\widehat{X}$ sobre $\widehat{X}$ en el `Paso 2` preservan el rango completo, es decir, $rango\,(\widehat{X}) = rango\,(\widehat{X} + \Delta\widehat{X}) = r$ y lo mismo para los errores regresivos de $\widehat{Y}$ en el `Paso 4`. Finalmente, la hipótesis técnica $\kappa_2(Y)\,n\,r^2\,\widetilde{\gamma}_n < 1$ se necesita para aplicar [43, Teorema 21.4] en el análisis de errores del `Paso 4`.

Presentaremos en el Teorema 3.5.2 dos enunciados para los errores regresivos del Algoritmo 3.5.1, uno con respecto a los factores calculados $\widehat{X}$, $\widehat{D}$ e $\widehat{Y}$ de A y otro con respecto a los factores exactos, el cual es el resultado usado generalmente. La razón para presentar estos dos enunciados es que el primero da errores regresivos por filas y columnas en $\widehat{X}$ e $\widehat{Y}$ más fuertes que el segundo. Esto puede ser usado para obtener errores regresivos más fuertes para algunas clases particulares de matrices, tales como matrices de Cauchy.

Teorema 3.5.2 *Sean $\widehat{X} \in \mathbb{R}^{m\times r}$, $\widehat{D} \in \mathbb{R}^{r\times r}$ e $\widehat{Y} \in \mathbb{R}^{r\times n}$ los factores de A calculados en el* `Paso 1` *del Algoritmo 3.5.1 y supongamos que satisfacen las cotas de error dadas en las ecuaciones* (2.4.1) *y* (2.4.2) *con respecto a los factores exactos X, D e Y de A. Supongamos también que* (2.4.3)*,*

$$\text{máx}\{\kappa_2(X), \kappa_2(Y)\}\,\sqrt{r}\,\widetilde{\gamma}_{r\,\text{máx}\{m,n\}} < 1 \qquad y \tag{3.5.3}$$

$$\kappa_2(Y)\,n\,r^2\,\widetilde{\gamma}_n < 1 \tag{3.5.4}$$

se cumplen. Sea $\widehat{x}_0$ la solución de mínima longitud calculada del problema de mínimos cuadrados $\text{mín}_{x\in\mathbb{R}^n} \|b - A\,x\|_2$ usando el Algoritmo 3.5.1 en precisión finita con unidad de redondeo `u`*. Entonces los siguientes enunciados se cumplen:*

(a) *$\widehat{x}_0$ es la solución de mínima longitud de*

$$\min_{x\in\mathbb{R}^n} \|(b + \Delta b) - (\widehat{X} + \Delta\widehat{X})(\widehat{D} + \Delta\widehat{D})(\widehat{Y} + \Delta\widehat{Y})\,x\|_2, \tag{3.5.5}$$

donde

$$\|\Delta\widehat{X}(:,j)\|_2 \le \widetilde{\gamma}_{mr}\,\|\widehat{X}(:,j)\|_2, \quad \|\Delta\widehat{Y}(j,:)\|_2 \le \widetilde{\gamma}_{nr}\,\|\widehat{Y}(j,:)\|_2, \quad \textit{para } j = 1,\ldots,r$$

$$|\Delta\widehat{D}| \le \widetilde{\gamma}_1\,|\widehat{D}|, \qquad \|\Delta b\|_2 \le \widetilde{\gamma}_{mr}\,\|b\|_2\,.$$

(b) $\widehat{x}_0$ *es la solución de mínima longitud de*

$$\min_{x\in\mathbb{R}^n} \|(b+\Delta b) - (X+\Delta X)(D+\Delta D)(Y+\Delta Y)\, x\|_2, \tag{3.5.6}$$

donde

$$\begin{aligned}
\|\Delta X\|_2 &\leq (\mathtt{u}\, p(m,n) + \sqrt{r}\,\widetilde{\gamma}_{mr} + \sqrt{r}\,\widetilde{\gamma}_{mr}\,\mathtt{u}\, p(m,n))\, \|X\|_2,\\
\|\Delta Y\|_2 &\leq (\mathtt{u}\, p(m,n) + \sqrt{r}\,\widetilde{\gamma}_{nr} + \sqrt{r}\,\widetilde{\gamma}_{nr}\,\mathtt{u}\, p(m,n))\, \|Y\|_2,\\
|\Delta D| &\leq (\mathtt{u}\, p(m,n) + \widetilde{\gamma}_1 + \widetilde{\gamma}_1\,\mathtt{u}\, p(m,n))\, |D|, \qquad \|\Delta b\|_2 \leq \widetilde{\gamma}_{mr}\, \|b\|_2\,.
\end{aligned}$$

(c) *Si* x_0 *es la solución de mínima longitud exacta de* $\min_{x\in\mathbb{R}^n} \|b - A\,x\|_2$*, entonces* $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ *puede acotarse como en el Teorema 3.4.1 con* $\alpha = (\mathtt{u}\, p(m,n) + \sqrt{r}\,\widetilde{\gamma}_{mr} + \sqrt{r}\,\widetilde{\gamma}_{mr}\,\mathtt{u}\, p(m,n))$*,* $\beta = (\mathtt{u}\, p(m,n) + \sqrt{r}\,\widetilde{\gamma}_{nr} + \sqrt{r}\,\widetilde{\gamma}_{nr}\,\mathtt{u}\, p(m,n))$*,* $\rho = (\mathtt{u}\, p(m,n) + \widetilde{\gamma}_1 + \widetilde{\gamma}_1\,\mathtt{u}\, p(m,n))$ *y* $\epsilon = \widetilde{\gamma}_{mr}$*. En particular, a primer orden en* $\mathtt{u}$*, y si* c *es una constante entera pequeña, entonces*

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq c\,\mathtt{u}\left[p_y(m,n)\,\kappa_2(Y) + p_x(m,n)\,\kappa_2(X)\,\frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2}\right] + O(\mathtt{u}^2)\,,$$

donde $p_y(m,n) := (p(m,n) + nr^{3/2})$ *y* $p_x(m,n) := (p(m,n) + mr^{3/2})$*.*

Demostración. Para demostrar la parte (a) escribamos los errores regresivos en los pasos 2, 3 y 4 del Algoritmo 3.5.1.

1. Los errores regresivos del `Paso 2` están dados en [43, Teorema 20.3]: la solución calculada en el `Paso 2`, $\widehat{x}_1$, es la solución exacta del problema de mínimos cuadrados

$$\min_{x\in\mathbb{R}^r} \|(b+\Delta b) - (\widehat{X}+\Delta\widehat{X})\, x\|_2\,, \tag{3.5.7}$$

donde $\|\Delta\widehat{X}(:,j)\|_2 \leq \widetilde{\gamma}_{mr}\|\widehat{X}(:,j)\|_2$, para $j = 1,\ldots,r$, y $\|\Delta b\|_2 \leq \widetilde{\gamma}_{mr}\,\|b\|_2$. Por lo tanto, $\|\Delta\widehat{X}\|_2 \leq \|\Delta\widehat{X}\|_F \leq \widetilde{\gamma}_{mr}\|\widehat{X}\|_F \leq \sqrt{r}\widetilde{\gamma}_{mr}\|\widehat{X}\|_2$. Note que, como mencionamos anteriormente, las ecuaciones (2.4.1) y (2.4.3) implican $rango\,(X) = rango\,(\widehat{X}) = r$, por lo tanto el teorema de perturbaciones de Weyl [69] para valores singulares y (3.5.3) implican $|\sigma_r(\widehat{X} + \Delta\widehat{X}) - \sigma_r(\widehat{X})|/\sigma_r(\widehat{X}) \leq \|\Delta\widehat{X}\|_2/\sigma_r(\widehat{X}) \leq \sqrt{r}\widetilde{\gamma}_{mr}\kappa_2(\widehat{X}) < 1$, y finalmente, $rango\,(\widehat{X}) = rango\,(\widehat{X} + \Delta\widehat{X}) = r$. Como consecuencia, $\widehat{x}_1$ satisface

$$\widehat{x}_1 = (\widehat{X} + \Delta\widehat{X})^\dagger(b + \Delta b), \tag{3.5.8}$$

con $\widehat{X} + \Delta\widehat{X} \in \mathbb{R}^{m\times r}$ tal que $rango\,(\widehat{X} + \Delta\widehat{X}) = r$.

2. Como consecuencia de [43, Lema 3.5], la solución, $\widehat{x}_2$, calculada en el `Paso 3` cumple que

$$(\widehat{D} + \Delta\widehat{D})\,\widehat{x}_2 = \widehat{x}_1 \quad \text{con} \quad |\Delta\widehat{D}| \leq \widetilde{\gamma}_1 |\widehat{D}|, \tag{3.5.9}$$

con $\widehat{D} + \Delta\widehat{D} \in \mathbb{R}^{r\times r}$ diagonal y no singular, dado que (2.4.2) y (2.4.3) implican $rango\,(D) = rango\,(\widehat{D}) = r$ y $\widetilde{\gamma}_1 < 1$ por (3.5.3).

3. Los errores regresivos del `Paso 4` están dados en [43, Teorema 21.4]. Para aplicar [43, Teorema 21.4] necesitamos que $rango\,(\widehat{Y}) = r$, el cual se sigue de (2.4.1) y (2.4.3), y la hipótesis

$$\| \, |\widehat{Y}^\dagger| \, |\widehat{Y}| \, \|_2 \, r \, n \, \gamma_n < 1,$$

se cumple por la ecuación (3.5.4), dado que $\| \, |\widehat{Y}^\dagger| \, |\widehat{Y}| \, \|_2 \, r \, n \, \gamma_n \leq \kappa_2(\widehat{Y})\, r^2 \, n \, \gamma_n < 1$. Con esta condición, la solución de mínima longitud calculada en el `Paso 4` $\widehat{x}_0$, es la solución mínima exacta del sistema lineal indeterminado

$$(\widehat{Y} + \Delta\widehat{Y})x = \widehat{x}_2\,,$$

con $\|\Delta\widehat{Y}(j,:)\|_2 \leq \widetilde{\gamma}_{nr}\|\widehat{Y}(j,:)\|_2$, para $j = 1, \ldots, r$. Además, podemos demostrar que $rango\,(\widehat{Y}) = rango\,(\widehat{Y} + \Delta\widehat{Y}) = r$ utilizando un argumento similar al usado para demostrar el resultado análogo para $\widehat{X} + \Delta\widehat{X}$. Por lo tanto, $\widehat{x}_0$ satisface

$$\widehat{x}_0 = (\widehat{Y} + \Delta\widehat{Y})^\dagger \widehat{x}_2, \tag{3.5.10}$$

con $\widehat{Y} + \Delta\widehat{Y} \in \mathbb{R}^{r\times n}$ y $rango\,(\widehat{Y} + \Delta\widehat{Y}) = r$.

De las ecuaciones (3.5.8), (3.5.9) y (3.5.10) tenemos que

$$\widehat{x}_0 = (\widehat{Y} + \Delta\widehat{Y})^\dagger (\widehat{D} + \Delta\widehat{D})^{-1} (\widehat{X} + \Delta\widehat{X})^\dagger (b + \Delta b) \tag{3.5.11}$$

$$= \left[(\widehat{X} + \Delta\widehat{X})\,(\widehat{D} + \Delta\widehat{D})\,(\widehat{Y} + \Delta\widehat{Y}) \right]^\dagger (b + \Delta b)\,, \tag{3.5.12}$$

donde la segunda igualdad se obtiene del Lema 3.1.2-(c). Esto y las cotas que hemos desarrollado para $\widehat{X}$, $\widehat{D}$ e $\widehat{Y}$ demuestran la parte (a) del Teorema 3.5.2.

La demostración del Teorema 3.5.2-(b) se obtiene fácilmente de la parte (a). Las ecuaciones (2.4.1) y (2.4.2) nos permiten escribir $\widehat{X} = X + E_X$, $\widehat{D} = D + E_D$ e $\widehat{Y} = Y + E_Y$, con $\|E_X\|_2 \leq \mathtt{u}\, p(m,n)\, \|X\|_2$, $|E_D| \leq \mathtt{u}\, p(m,n)|D|$ y $\|E_Y\|_2 \leq \mathtt{u}\, p(m,n)\|Y\|_2$. Por lo tanto, podemos escribir

$$\widehat{X} + \Delta\widehat{X} = X + E_X + \Delta\widehat{X} =: X + \Delta X\,, \tag{3.5.13}$$

donde

$$\begin{aligned}\|\Delta X\|_2 &\leq \|E_X\|_2 + \|\Delta \widehat{X}\|_2 \\ &\leq \mathtt{u}\, p(m,n)\, \|X\|_2 + \sqrt{r}\, \widetilde{\gamma}_{mr} \|\widehat{X}\|_2 \\ &\leq \mathtt{u}\, p(m,n)\, \|X\|_2 + \sqrt{r}\, \widetilde{\gamma}_{mr} \left(\|X\|_2 + \|E_X\|_2\right) \\ &\leq \left(\mathtt{u}\, p(m,n) + \sqrt{r}\, \widetilde{\gamma}_{mr} + \sqrt{r}\, \widetilde{\gamma}_{mr}\, \mathtt{u}\, p(m,n)\right) \|X\|_2 \,. \end{aligned} \tag{3.5.14}$$

Análogamente,

$$\begin{aligned}&\widehat{D} + \Delta \widehat{D} =: D + \Delta D\,, \quad \text{con } |\Delta D| \leq \left(\mathtt{u}\, p(m,n) + \widetilde{\gamma}_1 + \widetilde{\gamma}_1\, \mathtt{u}\, p(m,n)\right) |D|\,, \\ &\widehat{Y} + \Delta \widehat{Y} =: Y + \Delta Y\,, \quad \text{con } \|\Delta Y\|_2 \leq \left(\mathtt{u}\, p(m,n) + \sqrt{r}\, \widetilde{\gamma}_{nr} + \sqrt{r}\, \widetilde{\gamma}_{nr}\, \mathtt{u}\, p(m,n)\right) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \|Y\|_2\,. \end{aligned}$$

Si estas últimas ecuaciones junto con (3.5.13) y (3.5.14) se reemplazan en (3.5.5), entonces obtenemos la ecuación (3.5.6) y así demostramos la parte (b). Finalmente, la parte (c) es una consecuencia inmediata de (b) y el Teorema 3.4.1. □

Observemos que, como en una RRD los factores X e Y están bien condicionados, el Teorema 3.5.2-(c) garantiza que el error progresivo en la solución calculada por el Algoritmo 3.5.1 está acotado por $O(\mathtt{u})\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$.

3.6. Experimentos Numéricos

En esta sección mostraremos experimentos numéricos realizados usando MATLAB™ los cuales ilustran como son los errores cometidos por el Algoritmo 3.5.1 comparados con las predicciones teóricas y con los errores cometidos por el método usual para resolver problemas de mínimos cuadrados, es decir, usando la factorización QR calculada con el algoritmo tradicional de Householder implementado en MATLAB™ [43, Sección 20.2]. Para esto, usaremos tres clases importante de matrices rectangulares estructuradas que pueden tener números de condición muy grandes: Matrices de Cauchy, de Vandermonde y Graduadas. Para matrices pertenecientes a estas clases, RRDs precisas en el sentido de la Definición 2.4.2 pueden ser calculadas usando los algoritmos dados en [18] y [42]. Presentaremos sólamente experimentos para matrices $A \in \mathbb{R}^{m\times n}$ con entradas reales, $m \geq n$ y tales que $rango(A) = n$, lo cual significa que consideramos sólamente problemas de mínimos cuadrados con solución única.

Sabemos de (3.0.1) y (3.3.14) que si $\widehat{x}_0$ es la solución única de $\min_{x\in\mathbb{R}^n} \|Ax - b\|_2$ calculada por el algoritmo QR de MATLAB™ y x_0 es la solución exacta, entonces

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq c\, m\, n^{3/2}\, \mathtt{u}\, \kappa_2(A) \frac{\|A^\dagger\|_2 \|b\|_2}{\|x_0\|_2}, \tag{3.6.1}$$

la cual es una cota más grande que (3.0.1) pero confiable en la mayoría de situaciones. En cambio, el Algoritmo 3.5.1 satisface (ver (3.0.2) y el Teorema 3.5.2-(c))

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \mathtt{u}\, f(m,n)\ \left(\kappa_2(Y) + \kappa_2(X)\frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2}\right). \tag{3.6.2}$$

En nuestros experimentos, hemos calculado el error relativo en la solución para ambos algoritmos, y también las cantidades

$$\Theta_{QR} := \mathtt{u}\ \left(\kappa_2(A)\frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2}\right), \quad \Theta_{RRD} := \mathtt{u}\ \left(\kappa_2(Y) + \kappa_2(X)\frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2}\right), \tag{3.6.3}$$

con el fin de comprobar, la veracidad de las cotas (3.6.1) y (3.6.2).

Observamos que el Algoritmo 3.5.1, hasta ahora es el más preciso de los dos: para un vector aleatorio b, alcanza errores relativos en norma de orden la unidad de redondeo, lo que significa que Θ_{RRD} es $O(\mathtt{u})$ casi siempre, incluso para matrices A mal condicionadas.

3.6.1. Matrices de Cauchy

Las componentes de una matriz de Cauchy, $C \in \mathbb{R}^{m\times n}$, $m \geq n$, son definidas en términos de dos vectores $z = [z_1, \ldots, z_m]^T \in \mathbb{R}^m$ e $y = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$ como

$$c_{ij} = \frac{1}{z_i + y_j}, \qquad i = 1, \ldots, m,\ j = 1, \ldots, n. \tag{3.6.4}$$

Matrices de la forma $G = S_1 C S_2$, donde C es una matriz de Cauchy y, S_1 y S_2 son matrices diagonales y no singulares, en [18] son llamadas matrices quasi-Cauchy, las cuales incluyen, como un caso particular, las matrices de Cauchy para $S_1 = I_m$ y $S_2 = I_n$. Las matrices quasi-Cauchy tienen rango completo por columnas si $z_i \neq z_j$ para calquier $i \neq j$, $y_k \neq y_l$ para cualquier $k \neq l$ y $z_i \neq -y_j$ para todo i y j. El Algoritmo 3 en [18] utiliza una versión estructurada de GECP para calcular una RRD precisa de cualquier matriz quasi-Cauchy *cuadrada.* Este algoritmo puede ser extendido fácilmente para el caso de matrices rectangulares, y esta versión la usaremos en los experimentos de esta sección para calcular la RRD en el `Paso 1` del Algoritmo 3.5.1. El coste total de este paso es $2mn^2 - 2n^3/3 + O(n^2 + mn)$ operaciones más $mn^2/2 - n^3/6 + O(n^2 + mn)$ comparaciones.

Con el objetivo de hacer sencillas referencias, resumamos y mencionemos los dos algoritmos que son usados en esta sección para resolver $\min_{x\in\mathbb{R}^n} \|Cx - b\|_2$, donde C es una matriz de Cauchy:

- `LS-QR`: dados los vectores z e y, las componentes de C son calculadas como en (3.6.4) y el problema de mínimos cuadrados es resuelto usando la factorización QR de Householder implementada en `MATLAB`™.

- `LS-RRD`: el problema de mínimos cuadrados es resuelto usando el Algoritmo 3.5.1 y la RRD del `Paso 1` es calculada con la versión rectangular del Algoritmo 3 en [18] discutido anteriormente.

Las factorizaciones QR necesarias en los `Pasos 2` y `4` del Algoritmo 3.5.1 son calculadas con la rutina `qr` de `MATLAB`™. Observemos que en este caso, en el sistema lineal $Yx = x_2$ del `Paso 4` la matriz Y es no singular y por lo tanto GEPP se puede utilizar para obtener su solución.

En nuestros experimentos, hemos generado matrices de Cauchy con vectores z e y aleatorios, también hemos generado vectores b y hemos calculado la solución de $\text{mín}_{x\in\mathbb{R}^n} \|Cx - b\|_2$ usando los algoritmos `LS-QR` y `LS-RRD`. Para calcular los errores relativos $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$, tomamos como solución "exacta" x_0 la calculada utilizando el comando `svd` de `MATLAB`™ ejecutado en precisión aritmética variable. En cada experimento hemos fijado la precisión a $2\log_{10}|D_1/D_n| + 30$ dígitos decimales, donde D_1 y D_n son respectivamente, las entradas diagonales más grande y más pequeña (en valor absoluto) de la matriz diagonal D en la RRD de C calculada en el `Paso 1` del Algoritmo 3.5.1. La motivación de tomar $2\log_{10}|D_1/D_n| + 30$ dígitos decimales, viene del hecho de que $|D_1/D_n|$ tiene una magnitud similar a $\kappa_2(C)$, ya que X e Y están bien condicionadas, y esto, de acuerdo con (3.6.1) y a la discusión en la Sección 3.3.1, el error en el algoritmo tradicional para mínimos cuadrados es casi siempre mucho menor que $\mathtt{u}\,\kappa_2^2(C)$. Los vectores aleatorios z, y y b se han elegido de la distribución uniforme en el intervalo $[0, 1]$ (comando `rand` en `MATLAB`™), o bien, de la distribución normal (comando `randn` en `MATLAB`™). En todos los experimentos hemos analizado las ocho posibilidades resultantes en la selección de las distribuciones aleatorias para z, y y b.

Dos tipos de experimentos se han realizado. En el primer grupo hemos fijado el tamaño de la matriz: $m \times n = 100 \times 50, 50 \times 30$ ó 25×10. Para cada tamaño hemos generado 50×8 conjuntos diferentes de vectores aleatorios z, y y b, por lo tanto, generamos un total de 400 problemas de mínimos cuadrados diferentes para cada tamaño. En la Figura 3.6.1 mostramos los resultados para el caso 100×50 cuando los vectores z y b son seleccionados de la distribución normal y el vector y de la distribución uniforme en $[0, 1]$. Hemos graficado en una escala log-log el error relativo $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ de la solución, contra el número de condición de las matrices

(calculado a partir de la SVD "exacta" en precisión aritmética variable usada para calcular la solución "exacta" x_0) para los algoritmos `LS-QR` y `LS-RRD`. Aparte de esto, también hemos representado las cantidades Θ_{QR} y Θ_{RRD} que aparecen en la ecuación (3.6.3). Observamos que el error relativo en el algoritmo `LS-RRD` es de orden `u` veces una pequeña constante, como se predijo, mientras que el error para `LS-QR` es escalado casi lineal con respecto a $\kappa_2(C)$ hasta saturarse. La dependencia lineal en $\kappa_2(C)$ del error relativo en `LS-QR` es la predecida en la ecuación (3.6.1) dado que $\|C^\dagger\|_2\|b\|_2/\|x_0\|_2$ ha sido siempre moderada en estos experimentos. También podemos observar que la cota Θ_{RRD} es bastante buena y no sobreestima los errores actuales. Para otros tamaños y otras formas de generar z, y y b, los resultados han sido similares.

En nuestro segundo grupo de experimentos, hemos fijado el número de filas de la matriz y variado el número de las columnas. Hemos probado con matrices de tamaño $m = 100,\ n = 10:10:90$ (5×8 conjuntos de vectores aleatorios z, y y b para cada tamaño), $m = 50,\ n = 10:2:40$ (10×8 conjuntos de vectores aleatorios z, y y b para cada tamaño) y $m = 25,\ n = 5:5:20$ (20×8 conjuntos de vectores aleatorios z, y y b para cada tamaño). Esto hace un total de 2280 matrices. La Figura 3.6.2 muestra los resultados para $m = 50,\ n = 10:2:40$ para cuatro combinaciones diferentes de distribuciones aleatorias para z, y y b. Para cada tamaño graficamos el máximo error relativo de las 10 muestras. De nuevo, los errores relativos de la solución del algoritmo LS-RRD son de orden `u` veces una constante pequeña, mientras que para el algoritmo LS-QR son muy grandes. Para otros tamaños y otras formas de generar z, y y b se obtienen resultados similares.

Para todos los experimentos con matrices de Cauchy, el rango del número de condición ha sido $10^0 \lesssim \kappa_2(C) \lesssim 10^{100}$, el máximo valor del término $\|C^\dagger\|_2\|b\|_2/\|x_0\|_2$ ha sido 1376, $8 \leq \kappa_2(X) \leq 72$ y $13 \leq \kappa_2(Y) \leq 58$.

3.6.2. Matrices de Vandermonde

Hemos desarrollado experimentos numéricos del Algoritmo 3.5.1 similares a los presentados en la Sección 3.6.1 con matrices de Vandermonde. Las matrices de Vandermonde aparecen naturalmente en problemas de interpolación polinomial. Dado

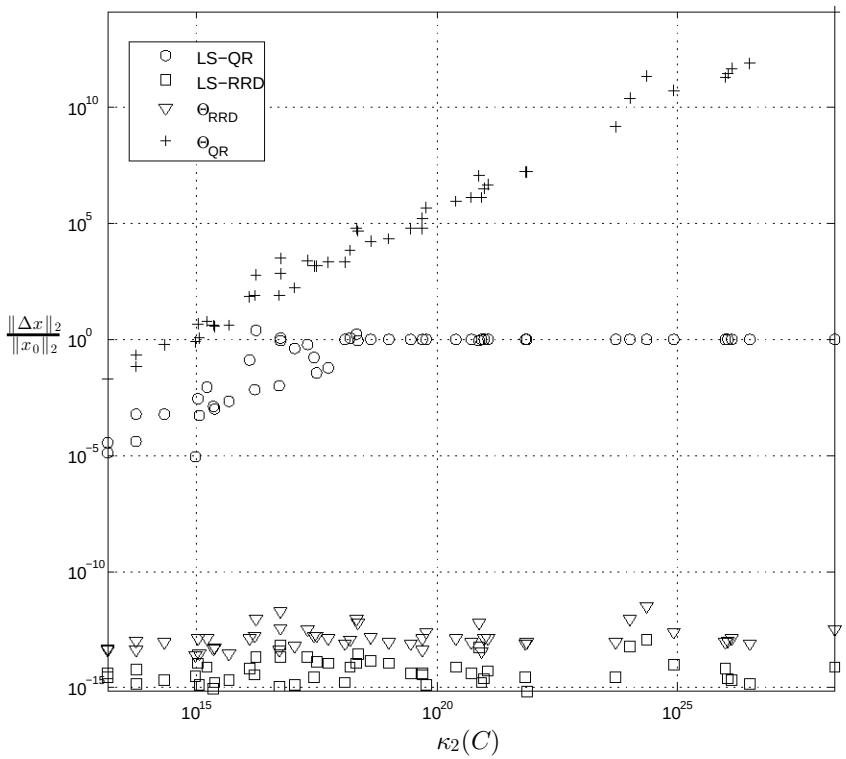

Figura 3.6.1: Error relativo progresivo $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ frente $\kappa_2(C)$. C matrices de Cauchy aleatorias de orden 100×50. Los vectores z y b son seleccionados de la distribución normal estándar y el vector y de la distribución uniforme en $[0, 1]$.

un vector de puntos $z = [z_1, \ldots, z_m]^T \in \mathbb{R}^m$ tal que $z_i \neq z_j$ si $i \neq j$ y un conjunto de "valores de la función" $b = [b_1, \ldots, b_m]^T \in \mathbb{R}^m$, el problema de encontrar el polinomio $P_{n-1}(z)$, de grado menor o igual a $n-1$, con $n \leq m$, que mejor se ajusta a los datos z y b en el sentido de mínimos cuadrados es equivalente a resolver el problema de mínimos cuadrados $\text{mín}_{c\in\mathbb{R}^n} \|Vc - b\|_2$ donde $V \in \mathbb{R}^{m\times n}$ es una matriz de Vandermonde, cuyas entradas están dadas por

$$v_{ij} = z_i^{j-1}, \quad i = 1, \ldots, m,\ j = 1, \ldots, n, \tag{3.6.5}$$

y la solución buscada $c \in \mathbb{R}^n$ es el vector que contiene los coeficientes del polinomio $P_{n-1}(z) = c_1 + c_2 z + \cdots + c_n z^{n-1}$. Un método que calcula una RRD precisa de cualquier matriz de Vandermonde fue presentado en [18, Sección 5] y está basado en el siguiente hecho, si $F \in \mathbb{R}^{n\times n}$ es la transformada discreta de Fourier de orden $n \times n$, entonces VF es una matriz quasi-Cauchy cuyos parámetros pueden ser calculados de forma precisa en $O(mn)$ operaciones, así como las sumas y restas de cualquier par de estos parámetros. Entonces, una RRD precisa de $VF = XDY$ puede ser calculada con el Algoritmo 3 en [18] adaptado al caso de matrices rectangulares como en la Sección 3.6.1. Finalmente, $V = X\,D\,(YF^T)$ es una RRD precisa de V. El coste total

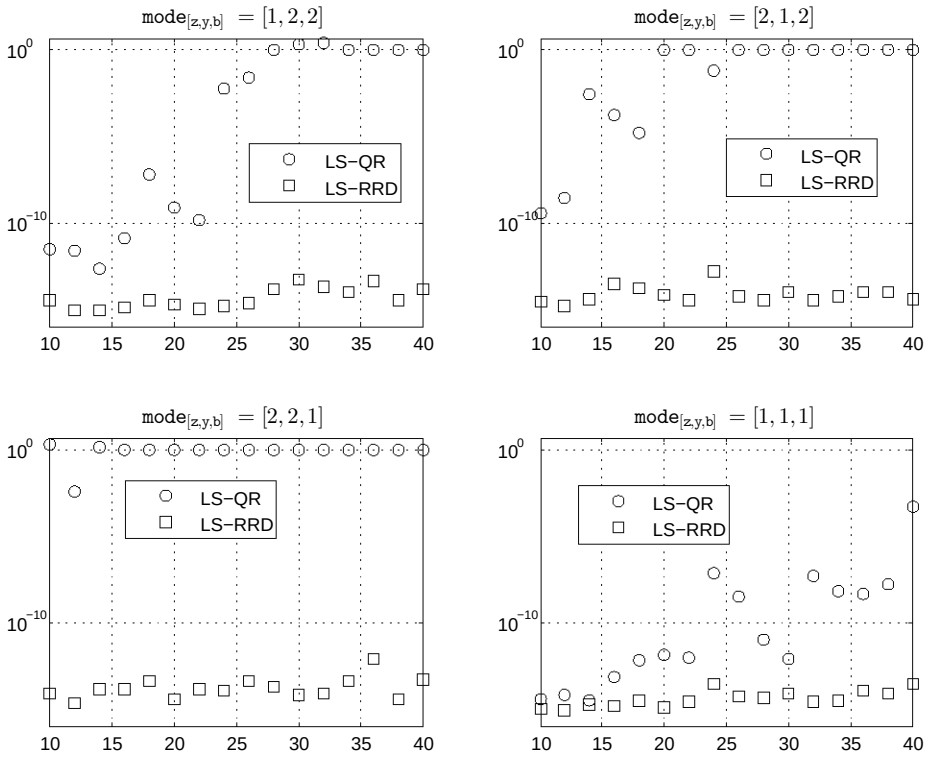

Figura 3.6.2: Error relativo progresivo $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ frente n, para matrices de Cauchy de orden $m \times n$ con $m = 50$ y $n = 10 : 2 : 40$ (10 matrices por cada tamaño) para cuatro combinaciones diferentes de distribuciones aleatorias (mode=1 denota la distribución normal estándar y mode=2 denota la distribución uniforme en $[0, 1]$) para z, y y b.

es el mismo que para matrices de Cauchy: $2mn^2 - 2n^3/3 + O(n^2 + mn)$ operaciones más $mn^2/2 - n^3/6 + O(n^2 + mn)$ comparaciones. Este algoritmo será usado para calcular la RRD en el `Paso 1` del Algoritmo 3.5.1. Los dos algoritmos usados en esta sección para resolver $\text{mín}_{x\in\mathbb{R}^n} \|Vx - b\|_2$ son:

- `LS-QR`: dado el vector z, las componentes de V son calculadas como en (3.6.5) y el problema de mínimos cuadrados es resuelto usando la factorización QR de Householder implementada en `MATLAB`™.
- `LS-RRD`: el problema de mínimos cuadrados es resuelto usando el Algoritmo 3.5.1 y la RRD del `Paso 1` es calculada con la versión rectangular discutida anteriormente del método presentado en [18, Sección 5].

En nuestros experimentos, generamos vectores aleatorios b y matrices de Vandermonde con vectores aleatorios z, y calculamos la solución de $\text{mín}_{x\in\mathbb{R}^n} \|Vx - b\|_2$ usando los algoritmos `LS-QR` y `LS-RRD`. Para calcular el error relativo $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$, seguimos el procedimiento presentado en la Sección 3.6.1 para obtener la solución

"exacta" x_0. Los vectores aleatorios z y b han sido elegidos de la distribución uniforme en $[0, 1]$ o bien, de la distribución normal. En todos los experimentos hemos probado las cuatro combinaciones resultantes de la elección de las distribuciones aleatorias para z y b.

Hemos probado con matrices $m \times n$ de tamaños $m = 50, n = 5 : 5 : 30$; $m = 100, n = 10 : 5 : 60$; y $m = 500, n = 100 : 50 : 250$. Para cada tamaño hemos generado diferentes conjuntos de vectores aleatorios z y b (para $m = 50, 100$, 25×4 diferentes conjuntos y para $m = 500$, 10×4 diferentes conjuntos), generando un total de 1860 problemas de mínimos cuadrados diferentes. La Figura 3.6.3 muestra los resultados para los tamaños $m = 100$ y $n = 10:5:60$, cuando los vectores z y b son seleccionados de la distribución normal. Hemos graficado en una escala log-log el error relativo $\|\widehat{x}_0 - x_0\|_2 / \|x_0\|_2$ de la solución frente el número de condición de las matrices (calculado a partir de la SVD "exacta" como en la Sección 3.6.1) para los algoritmos `LS-QR` y `LS-RRD`. Además, hemos graficado las cantidades Θ_{QR} y Θ_{RRD}[4] que aparecen en la ecuación (3.6.3). Los resultados son similares a los obtenidos en la Figura 3.6.1 para las matrices de Cauchy y referimos al lector a los comentarios hechos en la Sección 3.6.1. Para otros tamaños y otras formas de generar z y b los resultados han sido similares, produciendo el algoritmo `LS-RRD` errores relativos que son siempre de orden `u` veces una pequeña constante y el algoritmo `LS-QR` errores relativos escalados linealmente con respecto a $\kappa_2(V)$ y que son muy grandes para $\kappa_2(V)$ muy grandes. Para todos nuestros experimentos con matrices de Vandermonde, el rango de los números de condición ha sido $10^0 \lesssim \kappa_2(V) \lesssim 10^{70}$, el máximo valor del término $\|V^\dagger\|_2 \|b\|_2 / \|x_0\|_2$ ha sido 1076, $4 \leq \kappa_2(X) \leq 65$ y $3 \leq \kappa_2(Y) \leq 87$.

3.6.3. Matrices Graduadas

Otra clase de matrices para la cual es posible calcular una RRD precisa bajo ciertas condiciones son las matrices graduadas: matrices de la forma $A = S_1 B S_2 \in \mathbb{R}^{m \times n}$, con $S_1 \in \mathbb{R}^{m \times m}$ y $S_2 \in \mathbb{R}^{n \times n}$ matrices diagonales y no singulares que pueden ser arbitrariamente mal condicionadas, $B \in \mathbb{R}^{m \times n}$ una matriz bien condicionada, y $rango(A) = n$. Por lo tanto, la matriz A puede tener número de condición muy grande. Higham en [42] determina condiciones tales que si la factorización QR con

[4]Para mantener la escala de la figura graficamos $\min(\Theta_{QR}, 10)$ en vez de Θ_{QR}, dado que valores de Θ_{QR} mucho más grandes que los que aparecen en la Figura 3.6.1 han aparecido en este experimento.

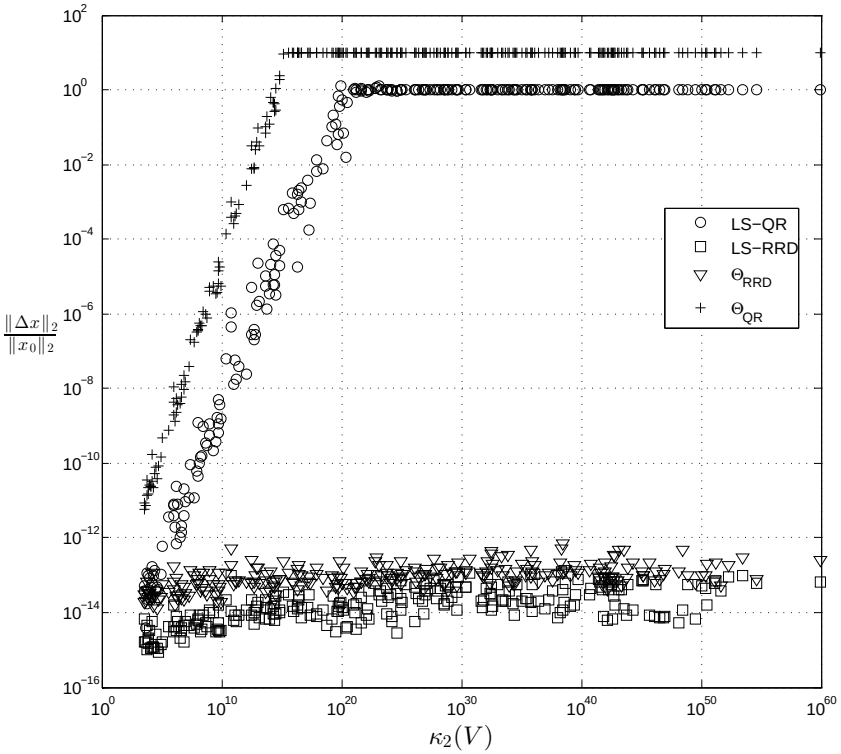

Figura 3.6.3: Error relativo progresivo $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ frente $\kappa_2(V)$ para matrices de Vandermonde aleatorias V de tamaños $100 \times 10{:}5{:}60$. Los vectores aleatorios z y b son seleccionados de la distribución normal estándar.

pivote completo (pivote por columnas junto con pivote por filas u ordenamiento por filas, para más detalles ver [42]) de una matriz graduada A es calculada y la SVD del factor R permutado es calculado por medio del Algoritmo 2.5.1, entonces la SVD de A se obtiene con alta precisión relativa. En esta sección mostramos que, bajo las mismas condiciones, la factorización QR con *pivote completo* puede usarse para resolver de manera precisa y eficiente el problema de mínimos cuadrados cuya matriz de coeficientes es graduada. Para este propósito, observemos que si $P_R A P_C = QR$ es la factorización QR reducida con pivote completo, donde P_R y P_C son matrices de permutaciones, $Q \in \mathbb{R}^{m\times n}$ y $R \in \mathbb{R}^{n\times n}$, entonces una RRD $A = XDY$ se puede obtener como se muestra a continuación:

$$A = P_R^T Q R P_C^T = (P_R^T\, Q)\, D\, (D^{-1}\, R\, P_C^T),$$

donde $D := diag(R)$, $X := P_R^T Q$ e $Y := D^{-1}\, R\, P_C^T$. En este caso, el `Paso 2` del Algoritmo 3.5.1 se reduce a $x_1 = Q^T P_R b$ y los `Pasos 3-4` pueden combinarse en uno sólo sin afectar los errores de redondeo, dado que los errrores son componente a componente y D es diagonal. Combinar estos pasos se hace simplemente para resolver el sistema lineal $(RP_C^T)x = x_1$. Por lo tanto, D no es necesario y el Algoritmo 3.5.1 se simplifica al Algoritmo 3.6.1, el cual es casi siempre el algoritmo usual para resolver

problemas de mínimos cuadrados usando la factorización QR.

Algoritmo 3.6.1 (Solución precisa para problemas de mínimos cuadrados vía QR con pivote completo cuando la matriz de coeficientes es graduada)

`Input:` $A \in \mathbb{R}^{m\times n}$ *matriz graduada,* $b \in \mathbb{R}^m$, $m \geq n = rango(A)$.

`Output:` x_0 solución única de $\min_{x\in\mathbb{R}^n} \|b - A\,x\|_2$

`Paso 1:` Calcular la descomposición QR reducida con pivote completo (ver [42]) de A, $A = P_R^T Q R P_C^T$.

`Paso 2:` Calcular $x_1 = Q^T\, P_R\, b$.

`Paso 3:` Resolver el sistema lineal consistente $R\, P_C^T\, x = x_1$ para obtener x_0.

El coste del Algoritmo 3.6.1 es esencialmente el mismo que el del método QR de Householder, es decir, $2mn^2 - \frac{2}{3}n^3$ flops, ya que el coste del pivotaje es $O(mn)$. Como es usual, no es necesario formar explícitamente la matriz Q. El Algoritmo 3.6.1 se puede aplicar también para matrices A que no sean graduadas, pero entonces la precisión no estaría garantizada.

Es importante notar que GECP también puede usarse para calcular bajo las mismas condiciones una RRD precisa de una matriz graduada [20, Sección 4] y el Algoritmo 3.5.1 puede usarse para resolver de manera precisa problemas de mínimos cuadrados cuando la matriz de coeficientes es graduada. Sin embargo, este procedimiento requiere calcular una factorización QR en el `Paso 2` del Algoritmo 3.5.1 y por lo tanto es más costoso que el Algoritmo 3.6.1.

Ahora, explicamos brevemente y de manera simplificada cuales son los errores cometidos por el Algoritmo 3.6.1. Estos errores son básicamente iguales a los del Algoritmo 3.5.1, si consideramos como RRD de A la siguiente factorización $A = (P_R^T\, Q)\, D\, (D^{-1}\, R\, P_C^T)$, y por lo tanto son determinados por los errores cometidos en el cálculo de la factorización QR con pivote completo. Para hacer la notación más simple, asumimos que la matriz A ha sido pre-pivoteada, es decir, que el `Paso 1` del Algoritmo 3.6.1 produce $P_R = I_m$ y $P_C = I_n$. Observe que esto induce correspondientemente pre-permutaciones en S_1, B y S_2. Primero, se demostró en [42, Teorema 2.5] que si $A = S_1 B S_2 \in \mathbb{R}^{m\times n} (m \geq n)$ con S_1 y S_2 matrices diagonales arbitrarias

y no singulares, entonces el factor $\widehat{R}$ calculado en el `Paso 1` del Algoritmo 3.6.1, satisface

$$S_1(B+\Delta B)S_2 = Q\widehat{R}, \qquad \text{con} \quad \|\Delta B\|_2 = O(\mathtt{u})\|B\|_2, \tag{3.6.6}$$

donde $Q \in \mathbb{R}^{m\times n}$ es una matriz con columnas ortonormales. La notación O-grande en (3.6.6) oculta algunos factores: factores de crecimiento y polinomios de grado menor en m y n, que pueden ser importantes en algunos casos. Para más detalles, ver [42]. Sin embargo, nuestros experimentos (como aquellos en [42]) muestran que en la práctica $O(\mathtt{u})$ es una constante pequeña multiplicada por $\mathtt{u}$. Si calculáramos el factor Q explícitamente, entonces obtendríamos una matriz $\widehat{Q}$ tal que $\|Q-\widehat{Q}\|_2 = O(\mathtt{u})$ [43, ecuación (19.13)], donde Q es la matriz que aparece en (3.6.6). Por lo tanto, en la siguiente discusión, no distinguiremos entre la Q exacta y la calculada.

Ahora necesitamos obtener, del error regresivo en (3.6.6), errores progresivos sobre la RRD del mismo tipo que aparece en la Definición 2.4.2. Con este fin, recordemos que si $B = LU$ es una factorización LU de B (sin pivote) con $L \in \mathbb{R}^{m\times n}$ y $U \in \mathbb{R}^{n\times n}$, cuyos factores están bien condicionados y tales que $\|B^\dagger\|_2 \approx \|L^\dagger\|_2\|U^{-1}\|_2$, entonces en [20, Teorema 4.1] se demostró que $S_1(B+\Delta B)S_2$ puede escribirse como[5]

$$S_1(B+\Delta B)S_2 = (I+E)A(I+F), \tag{3.6.7}$$

$$\text{con} \quad \max(\|E\|_2, \|F\|_2) = O(\tau\,\|\Delta B\|_2\|B^\dagger\|_2) = O(\mathtt{u})\,\tau\,\kappa_2(B), \tag{3.6.8}$$

donde el factor τ controla el gradiente (después de las permutacion en el `Paso 1` del Algoritmo 3.6.1) y está dado por

$$\tau = \max(1,\tau_1,\tau_2), \quad \text{con} \quad \tau_1 = \max_{\substack{1\le j\le n\\ j\le k\le m}} \frac{|(S_1)_{kk}|}{|(S_1)_{jj}|}, \;\; \tau_2 = \max_{1\le j\le k\le n} \frac{|(S_2)_{kk}|}{|(S_2)_{jj}|}. \tag{3.6.9}$$

Combinando (3.6.6) y (3.6.7), ignorando los términos de segundo orden y definiendo $\widehat{D} := diag(\widehat{R})$, obtenemos que $A = (I-E)Q\widehat{D}(\widehat{D}^{-1}\widehat{R})(I-F)$. Por lo tanto, $X = (I-E)Q$, $\widehat{D}$ e $Y = (\widehat{D}^{-1}\widehat{R})(I-F)$ pueden ser consideradas como los factores de una RRD exacta de A, cuyos factores calculados serían $\widehat{X} = Q$, $\widehat{D}$ e $\widehat{Y} = (\widehat{D}^{-1}\widehat{R})$. Observemos que en este caso los factores diagonales exactos y calculados son los mismos. Por lo tanto, podemos usar (3.6.8) para obtener $\max\{\|\widehat{X}-X\|_2/\|\widehat{X}\|_2\,,\, \|\widehat{Y}-Y\|_2/\|\widehat{Y}\|_2\} = O(\mathtt{u})\,\tau\,\kappa_2(B)$, lo cual nos permite aplicar el Teorema 3.5.2-(c) reemplazando $p(m,n)$

[5]Las actuales cotas son más complicadas e incluyen $\kappa_2(L)$ y $\kappa_2(U)$, y pueden ser mucho más grandes que $\kappa_2(B)$, ya que el pivotaje se realiza sobre A y no sobre B (para más detalle ver [20, Sección 4] y [42]). En nuestra discusión, pretendemos enfatizar los principales factores que controlan los errores en la práctica y no presentar un análisis riguroso.

por $\tau\,\kappa_2(B)$ y obtener a primer orden el siguiente error progresivo en la solución calculada por el Algoritmo 3.6.1

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq O(\mathtt{u})\,\tau\,\kappa_2(B)\,\left(\kappa_2(\widehat{D}^{-1}\widehat{R}) + \frac{\|A^{\dagger}\|_2\|b\|_2}{\|x_0\|_2}\right) + O(\mathtt{u}^2)\,, \tag{3.6.10}$$

donde usamos el hecho que $\kappa_2(Q) = 1$. Un punto clave en la cota (3.6.10) es el parámetro τ, el cual penaliza a $\kappa_2(B)$. Puede ser muy grande ya que las matrices diagonales S_1 y S_2 pueden ser arbitrariamente mal condicionadas. Sin embargo, las permutaciones provenientes de QR con pivote completo casi siempre reordenan S_1 y S_2 de forma que τ sea de orden uno.

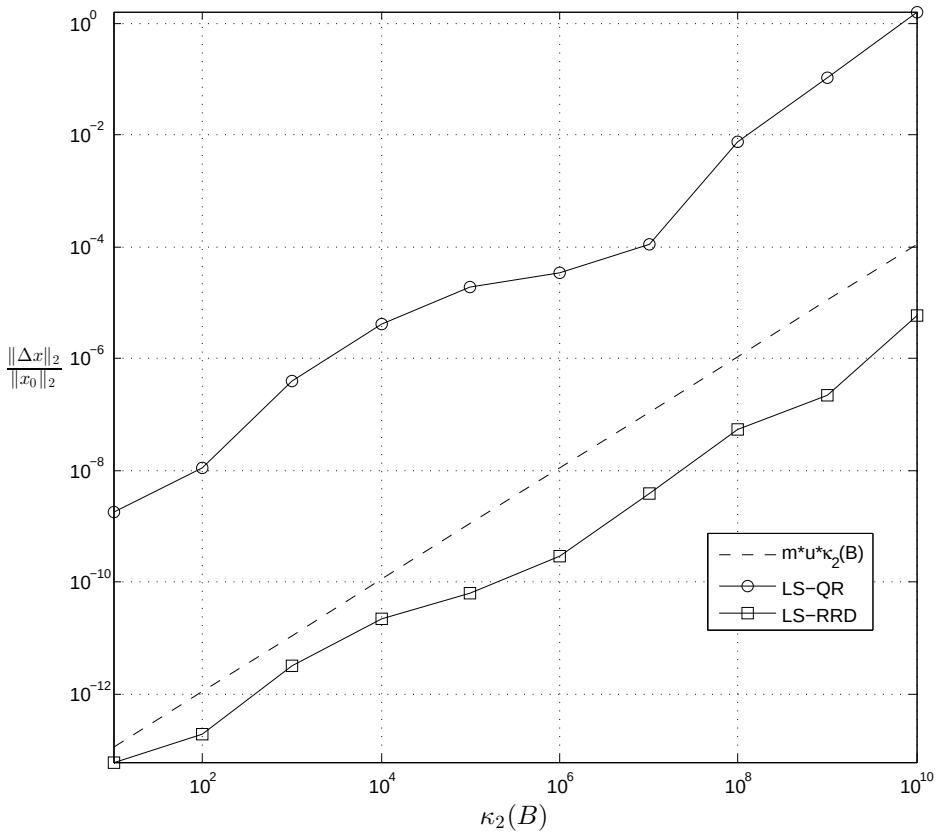

Figura 3.6.4: Error relativo progresivo $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ frente $\kappa_2(B)$ para matrices graduadas aleatorias $A = S_1 B S_2$ de tamaño 100×40, con B, S_1 y S_2 generados con la opción (b) que se explica en el texto.

Con el objetivo de comprobar la precisión del Algoritmo 3.6.1 para matrices graduadas A, hemos desarrollado diferentes experimentos numéricos similares a los presentados en [42] y hemos utilizado los siguientes dos Algoritmos para resolver $\min_{x\in\mathbb{R}^n} \|Ax - b\|_2$:

- `LS-QR`: calcula la solución usando la factorización QR de Householder con pivote por *columnas* implementada en `MATLAB`™.

- `LS-RRD`: usa el Algoritmo 3.6.1 y la factorización QR necesaria en el `Paso 1` se obtiene usando *ordenamiento por filas y pivotaje por columnas* [42].

Observe que los algoritmos `LS-QR` en las Secciones 3.6.1 y 3.6.2 usaron la factorización QR de Householder *sin* pivote. Ahora, usaremos pivote por columnas para ilustrar que el ordenamiento por filas adicional en `LS-RRD` es fundamental para obtener soluciones precisas. Para calcular el error relativo $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$, seguimos el procedimiento presentado en la Sección 3.6.1 con $D = diag(R)$.

En nuestros experimentos, hemos generado matrices aleatorias de la forma $A = S_1 B S_2$, donde B es construida siempre usando el modo 3 en la rutina `randsvd` de "Test Matrix Toolbox" desarrollado por Higham [41], es decir, B es una matriz densa aleatoria con un número de condición dado y con sus valores singulares distribuidos geométricamante. Las matrices diagonales S_1 y S_2 son generadas también con `randsvd` usando una variedad de distribuciones para valores singulares, es decir, para sus entradas diagonales. Hemos llevado a cabo experimentos donde los tamaños $m \times n$ de las matrices A y B han sido $m = 50,\ n = 10:10:30$ y $m = 100,\ n = 20:20:60$. Las matrices A se han construido de la siguiente manera: hemos seleccionado matrices B con números de condición $\kappa_2(B) = 10^i$, para $i = 1:10$. Hemos generado matrices diagonales S_1 y S_2 con entradas diagonales seleccionadas de uno de los tres pares de distribuciones: (a) las entradas de S_1 y S_2 tienen sus logarítmos uniformemente distribuidos (modo 5 en `randsvd`), pero las entradas de S_1 están ordenadas decrecientemente y las entradas de S_2 están ordenadas crecientemente; (b) las entradas de S_1 y S_2 están geométricamente distribuidas (modo 3 en `randsvd`) pero las entradas de S_1 están ordenadas crecientemente y las entradas de S_2 están ordenadas decrecientemente; (c) entradas distribuidas geométricamente en orden decreciente para S_1 y entradas con sus logarítmos uniformemente distribuidos ordenadas crecientemente para S_2. Tomamos $\kappa_2(S_1) = \kappa_2(S_2) = 10^k$ con $k = 2:2:16$. Para cada tamaño, para cada opción (a), (b) ó (c), y para cada tripleta $(\kappa_2(B), \kappa_2(S_1), \kappa_2(S_2))$, fueron generadas diez matrices y diez vectores b, cuyas entradas siguen la distribución normal. Esto hace, para cada tamaño y cada $\kappa_2(B)$, un total de $10 \times 8 \times 3 = 240$ matrices.

En la Figura 3.6.4 mostramos los resultados para matrices S_1 y S_2 de tamaño 100×40 con opción (b) para las matrices. Hemos graficado en una escala log-log el máximo error relativo para todas las 80 matrices con un $\kappa_2(B)$ en particular. Observamos que el error para el algoritmo `LS-RRD` se comporta como $O(\mathtt{u})\,\kappa_2(B)$, el cual,

de acuerdo con (3.6.10) implica que los números τ, $\kappa_2(\widehat{D}^{-1}\widehat{R})$ y $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ han sido de orden uno en todos estos experimentos. El error del algoritmo `LS-QR` pierde hasta seis dígitos de precisión, comportándose notablemente peor que `LS-RRD`. Hemos graficado una línea punteada mostrando la cantidad $m\,\mathtt{u}\,\kappa_2(B)$. El comportamiento para todos los otros tamaños de matrices y todos los otros modos de `randsvd` son similares.

Para todos nuestros experimentos con matrices graduadas el rango de los números de condición ha sido $10 \lesssim \kappa_2(A) \lesssim 10^{40}$ y el valor máximo del término $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ ha sido 108 y $2 \leq \kappa_2(\widehat{D}^{-1}\widehat{R}) \leq 18$.

3.6.4. Experimentos numéricos controlando el residuo

En los experimentos numéricos realizados anteriormente el vector b ha sido seleccionado aleatoriamente. Por lo tanto, el residuo relativo $\rho_r := \|b - Ax_0\|_2/\|b\|_2$ ha sido pocas veces muy pequeño, ya que si $rango(A) = n < m$, entonces es muy poco probable que $\theta(b, \mathcal{R}(A))$ sea muy pequeño. En esta sección, consideramos diferentes tipos de experimentos en los cuales generamos vectores aleatorios b con un valor fijo de ρ_r. Para este propósito, hemos desarrollado experimentos con matrices de Cauchy y Vandermonde de tamaño $m \times n$ de la siguiente manera. Primero obtenemos la RRD de la matriz $A = XDY$. Calculamos la SVD completa de la matriz X (lo cual se puede hacer de manera precisa utilizando el comando `svd` de `MATLAB`™ dado que X está bien condicionada): $X = U_X\Sigma_X V_X^T$, donde $U_X \in \mathbb{R}^{m\times m}$, $\Sigma_X \in \mathbb{R}^{m\times n}$ y $V_X \in \mathbb{R}^{n\times n}$. Entonces, particionamos $U_X = [U_1\ U_2]$, donde $U_1 \in \mathbb{R}^{m\times n}$ y $U_2 \in \mathbb{R}^{m\times(m-n)}$, y generamos vectores aleatorios $\alpha \in \mathbb{R}^n$ y $\beta \in \mathbb{R}^{m-n}$ tales que $\|\alpha\|_2 = \|\beta\|_2 = 1$. Con esto definimos

$$b_0 := U_1\alpha \in \mathcal{R}(A), \quad \Delta b := t\,U_2\beta \in \mathcal{R}(A)^\perp \quad \text{y} \quad b := b_0 + \Delta b\,, \tag{3.6.11}$$

donde hemos usado que $\mathcal{R}(A) = \mathcal{R}(X)$ y $t \geq 0$ es un parámetro. Observemos que de esta manera

$$\rho_r = \frac{\|b - Ax_0\|_2}{\|b\|_2} = \frac{t}{\sqrt{1+t^2}}, \tag{3.6.12}$$

ya que la solución x_0 satisface $Ax_0 = b_0$. Hemos usado matrices de Cauchy de tamaño $m = 100 \times n = 20 : 20 : 60$ y matrices de Vandermonde de tamaño $m = 50 \times n = 5 : 5 : 25$. Para cada tamaño hemos cambiado el valor de t para obtener residuos relativos $\rho_r = 10^{[-16:2:-2]}$. Para cada tamaño y para cada valor de ρ_r hemos generado 10 matrices y 10 vectores b, usando la distribución normal para

los parámetros de las matrices y para los vectores α y β. Finalmente, para cada valor de ρ_r, obtenemos el valor máximo de todos los errores relativos progresivos para todos los tamaños y todos los experimentos aleatorios. Los resultados para las matrices de Vandermonde son mostrados en la Tabla 3.6.1. Resultados similares fueron obtenidos para las matrices de Cauchy. Puede observarse que el análisis en la Sección 3.3.1 se cumple independientemente del tamaño del residuo relativo, aún cuando este es muy pequeño. Resaltamos nuevamente que el verdadero punto importante en la cota (3.0.2) es que $\|A^\dagger\|_2 \, \|b\|_2 / \|x_0\|_2$ es pequeño para casi todos los vectores b para cualquier tamaño fijo del residuo relativo no cercano a uno, independientemente del mal condicionamiento de la matriz A.

$\log_{10}(\|b - Ax_0\|_2 / \|b\|_2)$	-16	-14	-12	-10	-8	-6	-4	-2
`QR` : $\log_{10}(\|\Delta x\|_2 / \|x_0\|_2)$	$-2{,}7$	$-2{,}8$	$-2{,}1$	$-2{,}5$	$-3{,}8$	$-3{,}5$	$-3{,}3$	$-2{,}6$
`RRD` : $\log_{10}(\|\Delta x\|_2 / \|x_0\|_2)$	$-14{,}1$	$-14{,}0$	$-13{,}8$	$-13{,}9$	$-14{,}1$	$-14{,}0$	$-13{,}8$	$-13{,}8$

Tabla 3.6.1: Experimentos controlando el residuo. El error relativo progresivo $\|\Delta x\|_2 / \|x_0\|_2$ se muestra para ambos algoritmos `LS-QR` y `LS-RRD` descritos en la Sección 3.6.2 para diferentes valores del residuo relativo. Este experimento se hizo con matrices de Vandermonde de tamaños $m = 50 \times n = 5 : 5 : 25$. Todos los vectores aleatorios necesarios se eligieron de la distribución normal estándar.

3.6.5. Experimentos donde $\|A^\dagger\|_2 \, \|b\|_2 \, / \, \|x_0\|_2$ no es pequeño

En todos los experimentos aleatorios presentados en las Secciones 3.6.1, 3.6.2, 3.6.3 y 3.6.4 el factor $\|A^\dagger\|_2 \, \|b\|_2 / \|x_0\|_2$ ha sido moderado y por lo tanto el Algoritmo 3.5.1 resuelve de manera precisa todos los problemas ejecutados de mínimos cuadrados. Por lo tanto, estos experimentos han confirmado la discusión en la Sección 3.3.1. Por supuesto, es posible preparar experimentos donde $\|A^\dagger\|_2 \, \|b\|_2 / \|x_0\|_2$ no sea pequeño y el Algoritmo 3.5.1 no sea preciso, pero esto requiere seleccionar cuidadosamente los vectores b. Hemos procedido de la siguiente manera: Definimos los vectores b como $b = u_1 + b_\perp$, donde u_1 es el vector singular izquierdo de A correspondiente al valor singular más grande y $b_\perp$ es cualquier vector aleatorio ortogonal a $\mathcal{R}(A)$. Notemos que para vectores de este tipo $\|A^\dagger\|_2 \, \|b\|_2 / \|x_0\|_2 = \kappa_2(A) \sqrt{1 + \|b_\perp\|_2^2}$ y como consecuencia, los errores relativos progresivos en las soluciones cometidos por el Algoritmo 3.5.1 han sido grandes y proporcionales a la unidad de redondeo por el número de condición de las matrices. Sin embargo, sin importar este hecho, los errores del Algoritmo 3.5.1 han sido mucho más pequeños que los cometidos por el algoritmo QR de Householder. La razón es que el error (3.0.1) para la QR de Householder incluye el término Φ definido en (3.3.14) y para los vectores $b = u_1 + b_\perp$,

$\Phi = \kappa_2(A)^2\|b_\perp\|_2$, el cual es muy grande si $\kappa_2(A)$ es grande y $\|b_\perp\|_2 = \|r\|_2$ no es muy pequeño. Recordemos en este contexto nuestra discusión de la ecuación (3.3.15). Como ha sido resaltado en [32], note que para matrices mal condicionadas, los vectores $b = u_1 + b_\perp$ deben ser generados usando algoritmos altamente precisos para obtener u_1 y $b_\perp$ (ver [20] y Sección 3.6.4). Si el vector b es construido usando aritmética en coma flotante usual y el comando `svd` en MATLAB™, los errores de redondeo hacen imposible que b tenga exactamente la estructura requerida y el Algoritmo 3.5.1 calcule soluciones con alta precisión relativa.

4

Cálculo de autovalores y autovectores precisos de matrices simétricas graduadas

Cuando utilizamos algoritmos convencionales para calcular autovalores y autovectores de *matrices simétricas reales* mal condicionadas en aritmética en coma flotante, sólamente los autovalores con valor absoluto grande son calculados con precisión relativa garantizada. Los autovalores muy pequeños pueden calcularse sin precisión relativa, e incluso con el signo equivocado. Los autovectores están calculados con errores pequeños con respecto al gap relativo de los autovalores. Esto significa que si $\mathtt{u}$ es la unidad de redondeo y, q_i y $\widehat{q_i}$ son los respectivos autovectores exactos y calculados correspondientes al autovalor λ_i, entonces el ángulo agudo entre estos vectores está acotado por $\sin\theta(q_i, \widehat{q_i}) \leq O(\mathtt{u})/\mathrm{gap}_i$, donde $\mathrm{gap}_i = \mín_{j\neq i} |\lambda_i - \lambda_j| / \máx_k |\lambda_k|$. Lo cual implica que si existe más de un autovalor muy pequeño, entonces los correspondientes autovectores están calculados con errores grandes, incluso si los autovalores muy pequeños están bien separados en el sentido relatido. Ver [1, Sección 4.7] para un resumen sobre cotas de errores para problemas de autovalores simétricos.

Esto puede ser muy diferente si consideramos matrices con cierta estructura. Una matriz simétrica $A = A^T \in \mathbb{R}^{n\times n}$ es llamada graduada (o escalada) si $B := S^{-1}AS^{-1}$ es una matriz bien condicionada para alguna matriz diagonal $S = diag[s_1, \ldots, s_n]$, S será el escalamiento de la matriz graduada A. En [54], Martin, Reinsch y Wilkinson demostraron que una matriz graduada tendrá autovalores pequeños insensibles a pequeñas perturbaciones relativas en los elementos de la matriz.

En este capítulo estamos interesados en encontrar condiciones sobre las matrices S y B para obtener soluciones precisas del problema de autovalores cuando A es una matriz simétrica. Nuestro objetido es diseñar un algoritmo para calcular autovalores y autovectores de matrices simétricas graduadas de orden $n \times n$ con *alta precisión*

relativa, respetando la simetría del problema, con coste $O(n^3)$, es decir, aproximadamente el mismo coste de los algoritmos convencionales para matrices simétricas. Por alta precisión relativa en un problema de autovalores, queremos decir que los autovalores λ_i, los autovectores q_i y las correspondientes partes calculadas $\widehat{\lambda}_i$ y $\widehat{q}_i$, satisfacen:

$$|\widehat{\lambda}_i - \lambda_i| \leq O(\mathtt{u})|\lambda_i| \quad \text{y} \quad \sin\theta(q_i, \widehat{q}_i) \leq \frac{O(\mathtt{u})}{relgap_\lambda(A, \lambda_i)} \quad \text{para} \quad i = 1, \ldots, n. \tag{4.0.1}$$

Estas condiciones garantizan que los nuevos algoritmos calculan *todos* los autovalores, incluyendo los más pequeños con el signo y dígitos principales correctos. Sin embargo, los autovectores correspondientes a autovalores muy pequeños relativamente bien separados son calculados de manera precisa. En el caso de autovalores múltiples, o un clúster de autovalores muy cercanos en el sentido relativo, la cota dada anteriormente para $\sin\theta(q_i, \widehat{q}_i)$ es muy grande o tiende a infinito. En este caso, entendemos por alta precisión relativa que los senos de los ángulos canónicos entre los subespacios invariantes perturbados y sin pertubar correspondientes al clúster de autovalores acotados por $O(\mathtt{u})$ sobre el gap relativo entre los autovalores en el clúster y aquellos fuera de él [51]. Lo cual significa que el nuevo algoritmo calculará bases precisas de subespacios invariantes correspondientes a clústers de autovalores bien separados en el sentido relativo del resto de los autovalores.

Hasta ahora la posibilidad de calcular los autovalores y autovectores con alta precisión relativa de matrices simétricas bien escaladas ha estado restringida para algunas clases especiales de matrices. En [3] los autores demostraron que los autovalores y los autovectores de matrices simétricas diagonalmente dominantes bien escaladas pueden ser calculados de manera precisa. En [25] se estudiaron las matrices simétricas escaladas definidas positivas (ver también [55, 56]). Esto es un resultado importante ya que el problema de autovalores para matrices definidas positivas fue el primero en ser tratado con el objetivo de obtener alta precisión relativa utilizando una factorización de Cholesky y rotaciones de Jacobi.

Existen resultados parciales para el caso indefinido. En [67, 70], la clase de matrices para las cuales es posible resolver el problema fue ampliado significativamente, aunque sus resultados se basaron en las propiedades del factor polar definido positivo de A, que no son fáciles de obtener a partir de la matriz A. Por tanto, no existen resultados, fáciles de comprobar, sobre las condiciones que $S = diag[s_1, \ldots, s_n]$ y B tienen que satisfacer para obtener soluciones precisas del problema de autovalores de matrices simétricas generales $A = SBS$. En este capítulo vamos a demostrar

que, contrario a lo convencional, y al caso definido positivo, no es suficiente que B esté bien condicionada y que los elementos diagonales de la matriz escalada estén ordenados decrecientemente, después de la permutación producida por el pivotaje. Demostramos en el Corolario 4.2.4 que bajo perturbaciones de la matriz B de tamaño $\|\delta B\|_2 \leq \epsilon \|B\|_2$, el error relativo en los autovalores y en los autovectores, para $i = 1, \ldots, n$, está dado por

$$\left| \frac{\widetilde{\lambda}_i - \lambda_i}{\lambda_i} \right| \leq \epsilon \, \tau \, O\left(\kappa_2(B)\right) + O(\epsilon^2) \tag{4.0.2}$$

$$|\sin\theta(q_i, \widetilde{q}_i)| \leq \epsilon \, \tau \, \frac{O\left(\kappa_2(B)\right)}{relgap_\lambda(A, \lambda_i)} + O(\epsilon^2) \tag{4.0.3}$$

donde $\theta(q_i, \widetilde{q}_i)$ es el ángulo agudo entre q_i y $\widetilde{q}_i$. El factor τ controla el escalamiento y es definido como el máximo de los siguientes tres términos,

$$\tau = max\{1, \tau_L, \tau_D\}, \quad \tau_L = \max_{j<k} \frac{s_k}{s_j} \quad \text{y} \quad \tau_D = \max_{blocks,i} \max \left\{ \frac{s_{i+1}}{s_i}, \frac{s_i}{s_{i+1}} \right\}. \tag{4.0.4}$$

Si A es una matriz bien escalada en el sentido usual, con los elementos diagonales de S ordenados decrecientemente, entonces $\tau_L \leq 1$, pero el *nuevo* factor τ_D nos dice que esto no es suficiente para obtener buen acondicionamiento bajo perturbaciones de la forma $S(B + \delta B)S$. El factor τ_D proviene de la presencia de los bloques 2×2 en la factorización de Bunch & Parlett de $A = LDL^T$ (ver Sección 4.1). Si existe un bloque 2×2 en las posiciones i e $i + 1$ en la matriz diagonal por bloques D y si existe un "salto", ya sea aumentando o disminuyendo en los elementos diagonales de S, s_i y s_{i+1}, habrá un "condicionamiento efectivo"de tamaño τ_D que amplifica la perturbación de entrada de tamaño relativo ϵ. Observemos que esto es un nuevo fenómeno, que no aparece por ejemplo en las perturbaciones de valores singulares de matrices graduadas. En [20] se demostró que

Teorema 4.0.1 *[20, Teorema 4.1]*

$$\frac{|\sigma_i(\widetilde{A}) - \sigma_i(A)|}{|\sigma_i(A)|} \leq \max\{1, \overline{\tau}\}\{\kappa(L) + \kappa(U)\} \, \|L^{-1}\| \, \|U^{-1}\| \, \|\delta B\| + O(\|\delta B\|^2),$$

donde $A = S_1 B S_2$, $B = LU$ *(sin pivote) y* $\overline{\tau} = \max(\max_{i<j} \frac{S_{1,j}}{S_{1,i}}, \max_{i<j} \frac{S_{2,j}}{S_{2,i}})$.

Esto demuestra que, si L y U están bien condicionadas, entonces la descomposición en valores singulares de A y $\widetilde{A} = S_1(B + \delta B)S_2$ tienen garantizada alta precisión relativa

cuando $\|\delta B\|/\|B\|$ es pequeña y $\overline{\tau} \leq 1$. Pero el parámetro $\overline{\tau}$ en [20] es equivalente a τ_L en (4.0.4). Como los valores singulares son el módulo de los autovalores, comparando la ecuación (4.0.2) y el Teorema 4.0.1 podemos ver que la principal diferencia es el factor τ_D en el Teorema 4.0.1. Mostraremos en este capítulo que este factor es necesario y aparece en la sensitividad del problema de autovalores, y por lo tanto, se oculta y no se muestra explícitamente en los otros factores en la cota del Teorema 4.0.1.

El algoritmo que proponemos ofrece la precisión que merecen las matrices simétricas graduadas $A = SBS$ dados por los resultados de teoría de perturbaciones (4.0.2) y (4.0.3). La idea para construir este algoritmo es la misma utilizada para construir el Algoritmo 1.0.1, el concepto de *descomposición que revela el rango* mencionada en la Introducción. Para este propósito utilizaremos la factorizacición LDL^T por bloques de Bunch & Parlett con pivote completo (factorización BP-LDL^T). La factorización BP-LDL^T no proporciona en general una verdadera RRD, ya que no garantiza que el número de condición del factor L sea pequeño, sin embargo en la práctica si lo es. En cualquier caso, indicamos en todos los resultados que presentamos en este capítulo explícitamente la dependencia sobre los números de condición adecuados.

Nuestro camino para demostrar (4.0.2) y (4.0.3) será similar al presentado en la Sección 3.3. Demostraremos que nuestro algoritmo introduce pequeñas perturbaciones multiplicativas regresivas de la matriz, y entonces utilizaremos el hecho bien conocido que estas pequeñas perturbaciones multiplicativas producen pequeñas perturbaciones multiplicativas en los autovalores y autovectores, Teoremas 2.2.2 y 2.2.3.

El Capítulo está organizado de la siguiente manera: revisamos en la Sección 4.1 resultados básicos relacionados con la factorización LDL^T la cual obtenemos por medio del método de Bunch & Parlett con pivote completo. En la Sección 4.2 mostramos como pasar de perturbaciones aditivas de matrices graduadas a perturbaciones multiplicativas. En la Sección 4.3 presentamos el algoritmo para calcular autovalores y autovectores con alta precisión relativa vía una RRD y el correspondiente análisis de errores progresivo y regresivo. La precisión de este algoritmo es comprobada por medio de un experimentos numérico en la Sección 4.4. Por último presentamos un ejemplo numérico ilustrativo.

4.1. Matrices simétricas indefinidas y método de pivotaje diagonal

En esta sección revisamos las propiedades principales de la factorización simétrica LDL^T por bloques obtenida por el método de Bunch & Parlett con pivote completo (factorización BP-LD^T). La cual nos proporcionará, bajo ciertas condiciones, una RRD precisa para una matriz simétrica, necesaria si queremos garantizar la máxima precisión del algoritmo.

La factorización más utilizada para matrices simétricas [1, 39, 43] es la factorización LU por bloques:

$$PAP^T = L\,D\,L^T, \tag{4.1.1}$$

donde P es una matriz de permutaciones, L es triangular inferior unitaria y D es diagonal por bloques, con bloques de dimensión 1 ó 2. Los bloques 2×2 diagonales son matrices indefinidas simétricas, y los correspondientes bloques diagonales de L son matrices identidad de orden 2. Este método es llamado generalmente como *método de pivotaje diagonal* [43, Sección 11.1] y puede implementarse con pivote parcial, pivote completo o "rook pivoting". Estamos interesados en calcular una RRD simétrica, por lo tanto, nos enfocaremos en el *método de Bunch & Parlett con pivote completo* [8], el cuál en la práctica produce matrices L bien condicionadas. Notemos que una factorización de la forma LDL^T no es una RRD, ya que la matriz D no es diagonal. Para obtener una RRD, realizaremos una factorización espectral de cada uno de los bloques 2×2 de la matriz D, así, $D = V\Omega V^T$ con Ω diagonal y V ortogonal y diagonal por bloques como en D. Finalmente

$$PAP^T = L\,D\,L^T = (LV)\Omega(LV)^T := X\Omega X^T, \tag{4.1.2}$$

es una RRD simétrica. Este procedimiento ha sido esencialmente introducido en [65] para calcular una descomposición indefinida simétrica GJG^T, donde $J = \mathrm{diag}(\pm1)$. Notemos que la factorización por bloques GJG^T puede calcularse fácilmente a partir de $X\Omega X^T$ como:

$$PAP^T = L\,D\,L^T = X\Omega X^T = \left(L\,V\sqrt{|\Omega|}\right)\,J\,\left(\sqrt{|\Omega|}V^T\,L^T\right) := GJG^T. \tag{4.1.3}$$

Sin embargo, si $X\Omega X^T$ es calculada de forma precisa, entonces la factorización GJG^T por bloques también es calculada de manera precisa, y viceversa. Se ha demostrado también en [28] que la factorización BP-LDL^T por bloques es calculada de forma

precisa si y sólo si la factorización $X\Omega X^T$ es calculada de manera precisa. En el resto del capítulo usaremos las tres factorizaciones a medida que se vayan necesitando.

Para ser más específicos, describimos el método a continuación. Sea Π una matriz de permutaciones tal que

$$\Pi A \Pi^T = \begin{bmatrix} E & C^T \\ C & B \end{bmatrix}, \tag{4.1.4}$$

donde E es una matriz no singular 1×1 ó 2×2. La matriz de pivotaje E y la matriz de permutaciones Π se eligen comparando los números $\mu_0 = \max_{i,j} |a_{ij}| \equiv |a_{rs}|$ $(r \geq s)$ y $\mu_1 = \max_i |a_{ii}| \equiv |a_{pp}|$. Si $\mu_1 \geq \alpha\mu_0$, donde α es un parámetro $(0 < \alpha < 1)$, entonces $E = a_{pp}$, y si $\mu_1 < \alpha\mu_0$, entonces E tiene dimensión 2 y $E_{21} = a_{rs}$. El valor clásico para el parámetro es $\alpha = (1+\sqrt{17})/8$ $(\approx 0{,}64)$. Entonces podemos factorizar la ecuación (4.1.4):

$$\Pi A \Pi^T = \begin{bmatrix} I & 0 \\ CE^{-1} & I \end{bmatrix} \begin{bmatrix} E & 0 \\ 0 & B - CE^{-1}C^T \end{bmatrix} \begin{bmatrix} I & E^{-1}C^T \\ 0 & I \end{bmatrix}. \tag{4.1.5}$$

Supongamos que E es una matriz de dimensión 2×2, y sea $E = U\Lambda U^T$ su factorización espectral ortogonal calculada por el método de Jacobi [39, Sección 8.5]. Entonces

$$\Pi A \Pi^T = \begin{bmatrix} U & 0 \\ CU\Lambda^{-1} & I \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & B - CE^{-1}C^T \end{bmatrix} \begin{bmatrix} U^T & \Lambda^{-1}U^TC^T \\ 0 & I \end{bmatrix}. \tag{4.1.6}$$

El procedimiento se repite recursivamente sobre el complemento de Schur $B - CE^{-1}C^T$. Este método tiene un coste de $n^3/3$ operaciones más $O(n^2)$, que es el coste de la diagonalización ortogonal del bloque 2×2. Dado que utilizar pivote completo requiere la submatriz completa, la cual se busca en cada etapa, son necesarias hasta $n^3/6$ comparaciones.

El análisis de errores de esta factorización para cualquier estrategia de pivotaje es similar al de la factorización LU usual. La estabilidad de la factorización LDL^T por bloques está garantizada por el siguiente teorema.

Teorema 4.1.1 [43, Teorema 11.3] *Supongamos que la factorización LDL^T por bloques con cualquier estrategia de pivotaje es aplicada a una matriz simétrica $A \in \mathbb{R}^{n\times n}$ para obtener la factorización calculada $PAP^T \approx \widehat{L}\widehat{D}\widehat{L}^T$, donde P es una matriz de permutaciones y D tiene bloques diagonales de dimensión 1 ó 2, entonces*

$$P(A + \delta A)P^T = \widehat{L}\widehat{D}\widehat{L}^T \tag{4.1.7}$$

donde

$$|\delta A| \leq p(n)\mathtt{u}(|A| + P^T|\widehat{L}||\widehat{D}||\widehat{L}|^T P) + O(\mathtt{u}^2), \tag{4.1.8}$$

con $p(n)$ un polinomio lineal.

La estrategia en la elección de la permutación determina cómo acotar la norma de los factores $|L||D||L|^T$ en términos de $\|A\|$. El método de Bunch & Parlett con pivote completo es el más costoso pero determina las mejores cotas. Otros métodos para elegir permutaciones pueden encontrarse en [43, Secciones 11.1.2 y 11.1.3], pero una dificultad que tenemos con este método es que no se puede acotar el factor L de la factorización LDL^T calculada. Sin embargo, para el método de Bunch & Parlett [43, Teorema 8.12, Problema 8.5] se puede demostrar que

$$\|L\|_\infty \leq 2{,}78\, n. \tag{4.1.9}$$

y que

$$\kappa_\infty(L) \leq (3{,}78)^n\, n. \tag{4.1.10}$$

Esta cota es parecida a la obtenida para GECP. Por lo tanto, existe una remota posibilidad de que la estrategia de pivotaje de Bunch & Parlett falle al calcular el factor bien condicionado L.

El número de condición de la matriz diagonal por bloques D está acotado por el factor de crecimiento, definido como en la GE:

$$\rho_n = \frac{\text{máx}_{i,jk}\, |a_{i,j}^{(k)}|}{|a_{i,j}|}, \tag{4.1.11}$$

el cual involucra todos los elementos del complemento de Schur $a_{i,j}^{(k)}$ $(k = 1 : n)$ que aparecen durante el proceso. Para la factorización BP-LDL^T por bloques, el factor de crecimiento está acotado por $\rho \leq (1 + \alpha^{-1})^{n-1} = (2{,}57)^{n-1}$. Dicha cota es pesimista, sin embargo, en [7] el autor muestra que:

$$\rho \leq B_{cp}\, 3{,}07(n-1)^{0{,}446}, \tag{4.1.12}$$

donde B_{cp} es la cota para el factor de crecimiento de la factorización LU con pivote completo.

En resumen, no se puede garantizar que, en general, la factorización BP-LDL^T proporcione una verdadera RRD, con $\kappa_\infty(L)$ pequeño. En cualquier caso, indicamos

en todos los resultados que presentamos en este capítulo explícitamente la dependencia sobre los números de condición adecuados. Por conveniencia y claridad es necesario introducir alguna notación adicional relacionada con la factorización LDL^T que será utilizada en el resto del cápítulo.

Definición 4.1.2 *Dada una factorización $PAP^T = LDL^T$ de una matriz simétrica $A \in \mathbb{R}^{n\times n}$, como en la ecuación (4.1.1), donde D tiene un conjunto de bloques de dimensión 1×1 ó 2×2. Definimos la* **estructura de bloques** *asociada a la factorización LDL^T como el conjunto de enteros $I_{LDL^T} \subseteq \{1:n\}$ tales que $j \in I_{LDL^T}$ si un bloque 2×2 inicia en la posición (j,j) de la matriz D.*

Además, para cualquier matriz $M \in \mathbb{R}^{n\times n}$, dada una estructura de bloques I_{LDL^T}, definimos

1. *$M_L := stril_B(M)$ la parte triangular inferior estricta por bloques de M.*
2. *$M_D := diag_B(M)$ la parte diagonal por bloques de M.*
3. *$M_{[L+\frac{1}{2}D]} := M_L + \frac{1}{2}M_D$.*

A partir de las definiciones presentadas anteriormente, resulta sencillo demostrar algunas propiedades básicas:

Lema 4.1.3 *Sea $M \in \mathbb{R}^{n\times n}, I_{LDL^T}$ una estructura de bloques, y $C \in \mathbb{R}^{n\times n}$ una matriz diagonal por bloques con respecto a I_{LDL^T}. Entonces:*

$$(M\,C)_D = M_D\,C, \quad (C\,M)_D = C\,M_D, \tag{4.1.13}$$

$$(M\,C)_L = M_L\,C, \quad (C\,M)_L = C\,M_L. \tag{4.1.14}$$

4.2. Teoría de perturbaciones de matrices simétricas graduadas

En esta sección presentamos resultados para matrices simétricas reales graduadas $A = \underline{S}B\underline{S}$, donde $\underline{S}$ es diagonal y representa el escalamiento de A y B es una matriz que factorizamos en la forma $P^TBP = L_BD_BL_B^T$ como en la ecuación (4.1.1). El resultado principal en esta sección es el Teorema 4.2.3, que nos permite escribir perturbaciones aditivas de matrices graduadas como perurbaciones multiplicativas.

Entonces, usando los resultados de perturbaciones multiplicativas dados en los Teoremas 2.2.2 y 2.2.3, obtenemos en el Corolario 4.2.4, una cota para la sensibilidad del problema de autovalores cuando la matriz es simétrica graduada. Primero reproducimos de [31] dos lemas que serán necesarios en nuestro análisis. Estos dos lemas, nos proporcionan la sensibilidad a primer orden de los factores de la factorización por bloques LDL^T y GJG^T a perturbaciones aditivas de la matriz.

Lema 4.2.1 *[31, Teoremas 5.1 y 6.2] Sean $A \in \mathbb{R}^{n\times n}$ y $\widetilde{A} = A+E$ matrices simétricas y no singulares con factorización BP-LDL^T por bloques $PAP^T = LDL^T$ y $P\widetilde{A}P^T = \widetilde{L}\widetilde{D}\widetilde{L}^T$, respectivamente. Supongamos además que $\|L^{-1}E'L^{-T}D^{-1}\| < 1$ con $E' = PEP^T$. Entonces, a primer orden en $\|E'\|$,*

$$\begin{aligned} \widetilde{L} &= L + L(L^{-1}E'L^{-T}D^{-1})_L \quad y \\ \widetilde{D} &= D + (L^{-1}E'L^{-T})_D. \end{aligned}$$

En lo que resta del capítulo también necesitaremos la factorización GJG^T de una matriz de la forma $J + F$:

Lema 4.2.2 *Sea $J + F \in \mathbb{R}^{n\times n}$ una matriz simétrica y no singular, con $J = diag(\pm 1)$ una matriz diagonal, y supongamos que $\|F\| < 1/4$. Entonces, para cualquier estructura por bloques I_{LDL^T}, una factorizaciacion simétrica GJG^T de $J + F$, a primer orden en $\|F\|$ es,*

$$J + F = \overline{G}J\overline{G}^T \quad con \quad \overline{G} = I_n + F_L J + \frac{1}{2}F_D J + O(\|F\|^2).$$

Demostración. Escribimos

$$J + F = (I + F_L J + R_L J)(J + F_U + R_U) \tag{4.2.1}$$

(donde F_U representa la parte triangular superior por bloques, incluyendo la diagonal de la matriz F). Esto define R_L y R_U, los cuales son *residuos de segundo orden*. Para demostrar que R_L y R_U son términos de segundo orden en $\|F\|$, procederemos como en [68]. De la ecuación (4.2.1) tenemos una expresión para $R_L + R_U$

$$-(R_L + R_U) = -R = F_L J F_U + F_L J R_U + R_L J F_U + R_L J R_U. \tag{4.2.2}$$

Tomando normas y estudiando los residuos de segundo orden en $\|R\|$ tenemos

$$\|R\| \leq \frac{2\|F\|^2}{1 - 2\|F\| + \sqrt{1 - 4\|F\|}}. \tag{4.2.3}$$

Ahora, a primer orden en $\|F\|$

$$J + F = (I + F_L J)(J + F_U) = (I + F_L J)J(I + JF_D + JF_L^T), \tag{4.2.4}$$

y escribimos una expresión, a primer orden en $\|F\|$, para $J(I + JF_D + JF_L^T)$:

$$J(I+JF_D+JF_L^T) = J(I+JF_D)(I+JF_L^T) = (I+\frac{J}{2}F_D)J(I+\frac{J}{2}F_D)(I+JF_L^T). \tag{4.2.5}$$

Sustituyendo la ecuación (4.2.5) en (4.2.4) tenemos el resultado deseado. □

Ahora utilizaremos los Lemas 4.2.1 y 4.2.2 para demostrar el siguiente teorema, el cual, es el resultado principal en esta sección. Con este teorema, demostramos como perturbaciones aditivas de la matriz B se pueden convertir a perturbaciones multiplicativas de la matriz $A = \underline{S}B\underline{S}$.

Teorema 4.2.3 *Sea $A \in \mathbb{R}^{n\times n}$ una matriz simétrica graduada $A = \underline{S}B\underline{S}$ con $B = P^T L_B D_B L_B^T P$ la factorización BP-LDL^T de B, producida por cualquier estrategia de pivote, con estructura de bloques I_{LDL^T}, $S = P\underline{S}P^T = diag[s_1, \ldots, s_n]$ $(s_i > 0)$, y $\widetilde{A} = \underline{S}(B + \delta B)\underline{S}$ también simétrica. Entonces, a primer orden en δB, podemos escribir $P\widetilde{A}P^T$ como*

$$P\widetilde{A}P^T = (I_n + E)PAP^T(I_n + E)^T \tag{4.2.6}$$

donde

$$|E| \;\leq\; \tau\,|L_B|\,\left[|L_B^{-1}|\,|P\delta BP^T|\,|L_B^{-T}|\right]_{[L+\frac{1}{2}D]}\,|D_B^{-1}|\,|L_B^{-1}| \tag{4.2.7}$$

y

$$\tau = max\{1, \tau_L, \tau_D\}, \quad \tau_L = \max_{j<k} \frac{s_k}{s_j} \quad y \quad \tau_D = \max_{blocks,i} \max \left\{ \frac{s_{i+1}}{s_i}, \frac{s_i}{s_{i+1}} \right\} \tag{4.2.8}$$

Demostración. Primero escribimos $B = P^T L_B D_B L_B^T P$ en la forma $B = P^T GJG^T P$, como en la ecuación (4.1.1):

$$P\widetilde{A}P^T = P\underline{S}(P^T GJG^T P + \delta B)\underline{S}P^T = SG(J + G^{-1}\underline{\delta B}G^{-T})G^T S, \tag{4.2.9}$$

donde $\underline{\delta B} := P\delta BP^T$. La matriz $J + G^{-1}\underline{\delta B}G^{-T}$ es simétrica, por lo tanto usando el Lema 4.2.2 podemos calcular a primer orden en $F := G^{-1}\underline{\delta B}G^{-T}$, su factorización GJG^T:

$$J + F := J + G^{-1}\underline{\delta B}G^{-T} = \overline{G}J\overline{G}^T \tag{4.2.10}$$

donde, usando la notación presentada en el Lema 4.1.3,

$$\overline{G} = I_n + F_{[L+\frac{1}{2}D]}J. \tag{4.2.11}$$

Reemplazando la ecuación (4.2.10) en (4.2.9) y escribiendo $J = G^{-1}S^{-1}PAP^TS^{-1}G^{-T}$, tenemos que:

$$P\widetilde{A}P^T = SG\overline{G}J\overline{G}^TG^TS = (I_n + E)PAP^T(I_n + E)^T \tag{4.2.12}$$

donde

$$E = SG\overline{G}G^{-1}S^{-1} - I_n := SHS^{-1} \tag{4.2.13}$$

Por lo tanto, de las ecuaciones (4.2.11) y (4.1.3), escribimos la siguiente expresión para H

$$H = G\overline{G}G^{-1} - I_n = L_BV|\Omega|^{1/2}F_{[L+\frac{1}{2}D]}J|\Omega|^{-1/2}V^TL_B^{-1},$$

pero $V|\Omega|^{1/2}$, es una matriz diagonal por bloques, luego por el Lema 4.1.3, reescribimos H como mostramos a continuación

$$H = L_B[L_B^{-1}\underline{\delta B}L_B^{-T}]_{[L+\frac{1}{2}D]}V|\Omega|^{-1/2}J|\Omega|^{-1/2}V^TL_B^{-1}.$$

Finalmente, como $D^{-1} = V\,|\Omega|^{-1/2}\,J\,|\Omega|^{-1/2}\,V^T$, tenemos

$$H = L_B[L_B^{-1}\underline{\delta B}L_B^{-T}]_{[L+\frac{1}{2}D]}D^{-1}L_B^{-1}. \tag{4.2.14}$$

y sustituyendo en la ecuación (4.2.13):

$$E = SHS^{-1} = SL_B[L_B^{-1}\underline{\delta B}L_B^{-T}]_{[L+\frac{1}{2}D]}D^{-1}L_B^{-1}S^{-1}. \tag{4.2.15}$$

Los elementos de la matriz de pertubaciones multiplicativas E están dados por

$$E_{ij} = \frac{s_i}{s_j}H_{ij},$$

y

$$|E| \leq \max_{i,j}\left|\frac{s_i}{s_j}\right|\,|H|. \tag{4.2.16}$$

Demostraremos ahora que

$$\max_{i,j}\left|\frac{s_i}{s_j}\right| = \tau \tag{4.2.17}$$

con τ definido como en la ecuación (4.2.8). Las matrices H y E son triangular inferior por bloques. Los elementos E_{ij} y las parejas de índices (i,j), pueden separarse en tres conjuntos disjuntos:

- Las componentes diagonales: $i = j$. En este caso

$$\left|\frac{s_i}{s_j}\right| = 1. \tag{4.2.18}$$

- Los elementos de la parte triangular inferior estricta de la matriz: $i > j$. En este caso

$$\left|\frac{s_i}{s_j}\right| \leq \tau_L. \tag{4.2.19}$$

- Las componentes en uno de los bloques diagonales 2×2: $j = i+1$ ó $j = i-1$. En este caso

$$\left|\frac{s_i}{s_j}\right| \leq \tau_D. \tag{4.2.20}$$

De las ecuaciones (4.2.18)-(4.2.20) y de la definición de τ tenemos (4.2.17). Finalmente, de las ecuaciones (4.2.14)-(4.2.16) obtenemos (4.2.7). □

A partir del teorema anterior obtenemos el siguiente corolario.

Corolario 4.2.4 *Con las mismas hipótesis del Teorema 4.2.3, y supongamos además que* $\{\lambda_1, \ldots, \lambda_n\}$ *y* $\{\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n\}$ *son respectivamente los autovalores de* $A = \underline{S}B\underline{S}$ *y* $\widetilde{A} = \underline{S}(B + \delta B)\underline{S}$*, y que* $q_1, \ldots, q_n$ *y* $\widetilde{q}_1, \ldots, \widetilde{q}_n$ *son los correspondientes autovectores ortonormales, con* $\|\delta B\|_2 \leq \epsilon \|B\|_2$*, entonces, para* $i = 1, \ldots, n$ *se cumple que*

$$\left|\widetilde{\lambda}_i - \lambda_i\right| \leq 2\,\epsilon\,\tau\,\kappa_2^3(L_B)\,\kappa_2(D_B)|\lambda_i| + O(\epsilon^2) \tag{4.2.21}$$

$$\sin\theta(q_i, \widetilde{q}_i) \leq \epsilon\,\tau\,\kappa_2^3(L_B)\,\kappa_2(D_B)\left(1 + \frac{2}{relgap_{\widetilde{\lambda}}(A, \lambda_i)}\right) + O(\epsilon^2), \tag{4.2.22}$$

donde $\theta(q_i, \widetilde{q}_i)$ *es el ángulo agudo entre* q_i *y* $\widetilde{q}_i$*.*

Demostración. De (4.2.7) tenemos

$$\begin{aligned}
\|E\|_2 &\leq \|E\|_F \leq \tau\|H\|_F \leq \tau\|L_B\|_2 \,\|\left[L_B^{-1}\,P\delta B P^T\,L_B^{-T}\right]_{[L+\frac{1}{2}D]}\|_F\,\|D_B^{-1}\|_2\,\|L_B^{-1}\|_2 \\
&\leq \frac{\tau}{\sqrt{2}}\,\|L_B\|_2\,\|L_B^{-1}\|_2\|\delta B\|_2\|L_B^{-T}\|_F\|D_B^{-1}\|_2\,\|L_B^{-1}\|_2 \\
&\leq \epsilon\,\tau\sqrt{\frac{n}{2}}\|L_B\|_2\,\|L_B^{-1}\|_2\|L_B\|_2\|D_B\|_2\,\|L_B^T\|_2\|L_B^{-T}\|_2\|D_B^{-1}\|_2\,\|L_B^{-1}\|_2, \qquad (4.2.23)
\end{aligned}$$

donde hemos utilizado el hecho que $\|\delta B\|_2 \leq \epsilon\|B\|_2 \leq \epsilon\|L\|_2\|D\|_2\,\|L^T\|_2$. De la ecuación (4.2.23) y de los Teoremas 2.2.2, 2.2.3 y 4.2.3 obtenemos (4.2.21) y (4.2.22). □

4.3. Algoritmos y análisis de errores

En esta sección presentamos un algoritmo para calcular autovalores y autovectores de matrices simétricas graduadas con alta precisión relativa.

El método consiste básicamente en calcular una factorización LDL^T con el algoritmo de Bunch & Parlett descrito en la Sección 4.1, y con esto calcular una descomposición que revele el rango simétrica diagonalizando ortogonalmente los bloques 2×2 de la matriz D y finalmente aplicar el Algoritmo de Jacobi implícito (Algoritmo 2.5.6) para obtener soluciones precisas del problema de autovalores una vez que una descomposición que revele el rango precisa sea dada [30]. El análisis de errores de los dos primeros pasos del algoritmo son presentados respectivamente en el Teorema 4.3.2 (factorización LDL^T) y el Lema 4.3.3 (RRD simétrica usando rotaciones de Jacobi de los bloques 2×2 en D). Finalmente, en el Teorema 4.3.4 presentamos el análisis de errores completo del Algoritmo 4.3.1, el cual incluye el efecto del tercer paso: el Algoritmo de Jacobi Implícito.

A continuación presentamos el Algoritmo 4.3.1 que calcula la solución del problema de autovalores de matrices simétricas. Este método, así como el análisis de errores presentado en el resto de esta sección, garantizan alta precisión relativa en la solución del problema de autovalores de matrices simétricas graduadas. La alta precisión en este algoritmo no depende sólamente del escalamiento de la matriz, sino también de la relación entre el escalamiento y la posición de los bloques 2×2 de la factorización de Bunch & Parlett. Esta relación está controlada por el parámetro τ (ver Teorema 4.2.3) como explicaremos en la discusión presentada después del Teorema 4.3.4.

Algoritmo 4.3.1 (Solución precisa del problema de autovalores de matrices simétricas graduadas)

`Input:` $A \in \mathbb{R}^{n\times n}$ matriz simétrica.

`Output:` $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ y $U \in \mathbb{R}^{n\times n}$ matrices de autovalores y autovectores, respectivamente.

`Paso 1:` Calcular la factorización BP-LDL^T por bloques de $A = P^T LDL^T P$, tal que $D = diag(D_1, \ldots, D_p)$, donde $D_i \in \mathbb{R}^{s_i \times s_i}$ son bloques simétricos con $s_i = 1$ ó $s_i = 2$.

`Paso 2:` Convertir $A = P^T LDL^T P$ en una descomposición que revele el rango $A = X\Omega X^T$.

a. $D_i = V_i \Omega_i V_i^T$ [28, Lema A.5], tal que
   Si $s_i = 1$, entonces $V_i = 1$ y $\Omega_i = D_i \in \mathbb{R}$.
   Si $s_i = 2$, entonces $V_i \in \mathbb{R}^{2\times 2}$ es ortogonal y $\Omega_i \in \mathbb{R}^{2\times 2}$ es diagonal.
b. $V = diag(V_1, \ldots, V_p)$ y $\Omega = diag(\Omega_1, \ldots, \Omega_p)$.
c. $X = P^T L V$.

`Paso 3:` Aplicar el Algoritmo 2.5.6 para obtener U y Λ tal que $A = U\Lambda U^T$.

El coste total del Algoritmo 4.3.1 es el siguiente. La factorización BP-LDL^T tiene un coste de $n^3/3$ operaciones más $n^3/6$ comparaciones por el pivote completo. El coste de la diagonalización ortogonal de los bloques 2×2 es $O(n)$. El coste del Algoritmo de Jacobi Implícito es $9\, n^3\, N_{sw}$ donde N_{sw} es el número de barridos de Jacobi [30]. Para el algoritmo clásico de Jacobi [39, Sección 8.5] N_{sw} es proporcional a $\log(n)$ (ver también Secciones 8 y 9 de [30])

El análisis de errores del primer paso del Algoritmo 4.3.1 está dado por el siguiente teorema.

Teorema 4.3.2 *Sea $A \in \mathbb{R}^{n\times n}$ una matriz simétrica y no singular, y $\widehat{L}, \widehat{D}$ los factores calculados por el algoritmo de Bunch & Parlett con matriz de permutaciones P, en aritmética finita con unidad de redondeo* $\mathtt{u}$*, entonces*

$$\widehat{L}\widehat{D}\widehat{L}^T = (I_n + E_1)PAP^T(I_n + E_1)^T \tag{4.3.1}$$

donde

$$|E_1| \leq q(n)\,\mathtt{u}\,|\widehat{L}|\left[|\widehat{L}^{-1}||\widehat{L}||\widehat{D}||\widehat{L}^T||\widehat{L}^{-T}|\right]_{[L+\frac{1}{2}D]}|\widehat{D}^{-1}||\widehat{L}^{-1}| + O(\mathtt{u}^2), \tag{4.3.2}$$

con $q(n)$ un polinomio lineal en n. Además, si para alguna matriz diagonal y no singular $S = diag[s_1, \ldots, s_n]$ con $s_i > 0$, definimos $\widehat{L}_B := S^{-1}\widehat{L}S$, $\widehat{D}_B := S^{-1}\widehat{D}S^{-1}$, entonces

$$|E_1| \leq q(n)\,\mathtt{u}\,\tau\,|\widehat{L}_B|\left[|\widehat{L}_B^{-1}||\widehat{L}_B||\widehat{D}_B||\widehat{L}_B^T||\widehat{L}_B^{-T}|\right]_{[L+\frac{1}{2}D]}|\widehat{D}_B^{-1}||\widehat{L}_B^{-1}| + O(\mathtt{u}^2), \tag{4.3.3}$$

donde

$$\tau := max\{1, \tau_L, \tau_D\}, \quad \tau_L := \max_{j<k}\frac{s_k}{s_j}, \quad y \quad \tau_D := \max_{blocks,i}\max\left\{\frac{s_{i+1}}{s_i}, \frac{s_i}{s_{i+1}}\right\}. \tag{4.3.4}$$

Demostración. Del Teorema 4.1.1 tenemos que:

$$A = P^T \widehat{L}\widehat{D}\widehat{L}^T P - \delta A \tag{4.3.5}$$

donde

$$|\delta A| \leq 2\, p(n)\, \mathtt{u}\, P^T |\widehat{L}||\widehat{D}||\widehat{L}^T| P + O(\mathtt{u}^2). \tag{4.3.6}$$

Aplicamos el Teorema 4.2.3 a la ecuación (4.3.5) teniendo en cuenta las siguientes equivalencias

Teorema 4.2.3	Ecuación (4.3.5)
$S = \underline{S}$	I_n
τ	$\tau = 1$
$A = B$	$P^T \widehat{L}\widehat{D}\widehat{L}^T P$
$\widetilde{A} = B + \delta B$	A
δB	$-\delta A$
$L_B D_B L_B^T$	$\widehat{L}\widehat{D}\widehat{L}^T$

con lo que podemos escribir la ecuación (4.3.5) como

$$PAP^T = (I_n - E_1)\widehat{L}\widehat{D}\widehat{L}^T (I_n - E_1)^T$$

donde

$$|E_1| \leq |\widehat{L}| \left[|\widehat{L}^{-1}|\, P\, |\delta A|\, P^T\, |\widehat{L}^{-T}| \right]_{[L+\frac{1}{2}D]} |\widehat{D}^{-1}||\widehat{L}^{-1}|. \tag{4.3.7}$$

Reemplazando la ecuación (4.3.6) en (4.3.7) tenemos:

$$|E_1| \leq 2\, p(n)\, \mathtt{u}\, |\widehat{L}| \left[|\widehat{L}^{-1}||\widehat{L}||\widehat{D}||\widehat{L}^T||\widehat{L}^{-T}| \right]_{[L+\frac{1}{2}D]} |\widehat{D}^{-1}||\widehat{L}^{-1}| + O(\mathtt{u}^2),$$

esto es (4.3.2).

Ahora, si $\widehat{L} = S\widehat{L}_B S^{-1}$ y $\widehat{D} = S\widehat{D}_B S$, entonces

$$|E_1| \leq 2\, p(n)\, \mathtt{u}\, S|\widehat{L}_B| \left[|\widehat{L}_B^{-1}||\widehat{L}_B||\widehat{D}_B||\widehat{L}_B^T||\widehat{L}_B^{-T}| \right]_{[L+\frac{1}{2}D]} |\widehat{D}_B^{-1}||\widehat{L}_B^{-1}|S^{-1} + O(\mathtt{u}^2). \tag{4.3.8}$$

La matriz que aparece al lado derecho de la desigualdad es triangular inferior por bloques como la matriz E del final de la demostración del Teorema 4.2.3, por lo tanto, procediendo de la misma manera tenemos (4.3.3). □

Presentamos una observación importante sobre la relación entre la matriz "original" de escalamiento $\underline{S}$, la matriz de permutaciones P, la matriz "final" de escalamiento S y la matriz "interior" B. La matriz de permutaciones P está fijada por medio de la factorización BP-LDL^T

$$A = \underline{S}B\underline{S} \approx P^T\widehat{L}\widehat{D}\widehat{L}^T P.$$

Suponemos que la matriz A está bien escalada por la matriz $\underline{S}$, pero en general, los elementos diagonales de $\underline{S}$ no están ordenados decrecientemente. El pivote completo tiende a poner los elementos más grandes en la parte superior, de tal manera que la matriz de escalamiento $S = P\underline{S}P^T$ de PAP^T tenga sus elementos diagonales ordenados decrecientemente,

$$PAP^T \approx S\widehat{L}_B S^{-1}\; S\widehat{D}_B S\; S^{-1}\widehat{L}_B^T S = S\widehat{L}_B\widehat{D}_B\widehat{L}_B^T S.$$

De esta manera, el producto $\widehat{L}_B\widehat{D}_B\widehat{L}_B^T$ desempeña la doble función de la matriz B y su factorización *exacta* $\widehat{L}_B\widehat{D}_B\widehat{L}_B^T$ (con la estrategia de pivotaje inducida por el pivote completo sobre A, que, en general, no será la permutación que produciría la estrategia de Bunch&Parlett sobre B) que aparecen en el Teorema 4.2.3. Claramente, ni la matriz B, ni las matrices $\widehat{L}_B$ ó $\widehat{D}_B$ son creadas en el algoritmo.

El análisis de errores del segundo paso del Algoritmo 4.3.1 lo presentamos en el siguiente Lema.

Lema 4.3.3 *Supongamos que el Algoritmo 4.3.1 es aplicado en un ordenador con unidad de redondeo* $\mathtt{u}$. *Sean* $\widehat{L}$ *y* $\widehat{D}$ *los factores y* P *la matriz de permutaciones calculados por la factorización BP-*LDL^T *en el primer paso, y* $\widehat{X}$ *y* $\widehat{\Omega}$ *los factores calculados en el segundo paso del Algoritmo 4.3.1. Además, supongamos que los bloques* 2×2 *de la matriz* $\widehat{D}$ *se han diagonalizado usando rotaciones de Jacobi como aparece en [28, Apéndice A.1]. Entonces*

$$\widehat{X}\widehat{\Omega}\widehat{X}^T = (I_n + E_2)\, P^T\widehat{L}\widehat{D}\widehat{L}^T P(I_n + E_2)^T \tag{4.3.9}$$

donde

$$\|E_2\|_2 \leq \left(c\,\sqrt{n}\mathtt{u} + O(\mathtt{u}^2)\right)\,\kappa_2(\widehat{L}), \tag{4.3.10}$$

con c *una pequeña constante entera.*

Demostración. Como los bloques de la matriz $\widehat{D}$ se han obtenido a partir de la factorización BP-LDL^T, el Lema A.5 de [28] se puede aplicar (con $k = 0$) para

demostrar que los factores $\widehat{V}$ y $\widehat{\Omega}$ calculados en el paso 2.b del Algoritmo 4.3.1 satisfacen, para $i = 1, \ldots, p$, (p es el número total de bloques de la matriz $\widehat{D}$),

$$\|\widehat{V}_i - V_i\|_F \leq c\,\mathtt{u} + O(\mathtt{u}^2) \tag{4.3.11}$$

$$|\widehat{\Omega}_i - \Omega_i| \leq \left(c\,\mathtt{u} + O(\mathtt{u}^2)\right) |\Omega|_i, \tag{4.3.12}$$

donde $\Omega = diag(\Omega_1, \ldots, \Omega_p)$ y $V = diag(V_1, \ldots, V_p)$ son los factores exactos. El factor $\widehat{X}$ es calculado como $\widehat{X} = \mathtt{f}l(P^T\widehat{L}\widehat{V})$. Como esta operación sólamente involura productos 2×2 de matrices ortogonales de la ecuación (4.3.11), tenemos

$$\|\widehat{X} - X\|_F \leq \left(c\,\mathtt{u} + O(\mathtt{u}^2)\right) \|\widehat{X}\|_F \tag{4.3.13}$$

donde X es el factor exacto de $P^T\widehat{L}\widehat{D}\widehat{L}^TP$.

Ahora, tenemos que

$$P^T\widehat{L}\widehat{D}\widehat{L}^TP = X\Omega X^T = (\widehat{X} + \delta X)(\widehat{\Omega} + \delta\Omega)(\widehat{X} + \delta X)^T. \tag{4.3.14}$$

De la ecuación (4.3.12)

$$\widehat{\Omega}_{ii} + \delta\Omega_{ii} = \widehat{\Omega}_{ii}(1 + \alpha_i) = \sqrt{1+\alpha_i}\ \widehat{\Omega}_{ii}\ \sqrt{1+\alpha_i} \quad \text{con} \quad |\alpha_i| \leq c\,\mathtt{u}$$

de

$$\widehat{\Omega} + \delta\Omega = (I_n + E_D)\widehat{\Omega}(I_n + E_D), \tag{4.3.15}$$

donde E_D diagonal y tal que $\|E_D\|_2 \leq c\,\mathtt{u}$ y de la ecaución (4.3.13) tenemos

$$\|\delta X\|_2 \leq \|\delta X\|_F \leq \left(c\,\mathtt{u} + O(\mathtt{u}^2)\right) \|\widehat{X}\|_F \leq \left(c\,\mathtt{u}\sqrt{n} + O(\mathtt{u}^2)\right) \|\widehat{X}\|_2.$$

Finalmente, reemplazando (4.3.15) en (4.3.14) obtenemos

$$\begin{aligned} P^T\widehat{L}\widehat{D}\widehat{L}^TP &= (I_n + \delta X\widehat{X}^{-1})\widehat{X}(I_n + E_D)\widehat{\Omega}(I_n + E_D)\widehat{X}^T(I_n + \delta X\widehat{X}^{-1})^T \\ &:= (I_n - E_2)\widehat{X}\widehat{\Omega}\widehat{X}^T(I_n - E_2)^T, \end{aligned}$$

donde

$$I_n - E_2 := (I_n + \delta X\widehat{X}^{-1})(I_n + \widehat{X}E_D\widehat{X}^{-1})$$

y

$$\|E_2\|_2 \leq \left(c\,\mathtt{u}\sqrt{n} + O(\mathtt{u}^2)\right) \kappa_2(\widehat{X}) \approx \left(c\,\mathtt{u}\sqrt{n} + O(\mathtt{u}^2)\right) \kappa_2(\widehat{L}),$$

ya que $\kappa_2(\widehat{X}) \approx \kappa_2(\widehat{L})$. □

Ahora combinamos el Teorema 4.3.2, el Lema 4.3.3 y el Teorema 6 de [30] para obtener el resultado final de errores multiplicativos regresivos de los autovalores y autovectores calculados por el Algoritmo 4.3.1.

Teorema 4.3.4 *Sea $A \in \mathbb{R}^{n\times n}$ una matriz simétrica y no singular, sean $\widehat{\Lambda}$ y $\widehat{U}$ las matrices de autovalores y autovectores de A calculados por el Algoritmo 4.3.1 en un ordenador con unidad de redondeo $\mathtt{u}$ y sean $\widehat{L}$ y $\widehat{D}$ los factores calculados por la factorización BP-LDL^T. Entonces existe una matriz ortogonal $U \in \mathbb{R}^{n\times n}$ tal que $\|\widehat{U} - U\|_F \leq c\, n^2\, \mathtt{u}$ y*

$$U\widehat{\Lambda}\, U^T = (I_n + E)\, A\, (I_n + E)^T \tag{4.3.16}$$

donde

$$\|E\|_2 \leq q(n)\,\mathtt{u}\left(\left\| |\widehat{L}| \left[|\widehat{L}^{-1}||\widehat{L}||\widehat{D}||\widehat{L}^T||\widehat{L}^{-T}| \right]_{L+\frac{1}{2}D} |\widehat{D}^{-1}||\widehat{L}^{-1}| \right\|_2 + \kappa_2(\widehat{L}) \right) + O(\mathtt{u}^2), \tag{4.3.17}$$

con $q(n)$ un polinomio cuadrático en n.

Demostración. Procederemos a primer orden en $\mathtt{u}$. Para obtener (4.3.16) sólamente tenemos que aplicar en orden inverso el análisis de errores del Algoritmo 4.3.1. En el último paso del algoritmo la descomposición espectral se obtuvo a partir de una RRD. Por lo tanto del Teorema 6 en [30], sabemos que existe una matriz ortogonal $U \in \mathbb{R}^{n\times n}$ tal que

$$U\widehat{\Lambda}\, U^T = (I_n + E_3)\, \widehat{X}\widehat{\Omega}\widehat{X}^T (I_n + E_3)^T$$

con $\|\widehat{U} - U\|_F \leq c\, n^2\, \mathtt{u}$ y

$$\|E_3\|_F \leq c\, n^2\, \mathtt{u}\, \kappa_2(\widehat{L}). \tag{4.3.18}$$

Luego, por el Teorema 4.3.2 y el Lema 4.3.3 tenemos los errores cometidos por los pasos 1 y 2; escribiendo todos estos resultados juntos tenemos

$$U\widehat{\Lambda}\, U^T = (I_n + E_3)(I_n + E_2)(I_n + P^T E_1 P)\, A\, (I_n + P^T E_1 P)^T (I_n + E_2)^T (I_n + E_3)^T, \tag{4.3.19}$$

con

$$\|E_2\|_2 \leq c\, \mathtt{u}\sqrt{n}\kappa_2(\widehat{L}) \tag{4.3.20}$$

y

$$\||E_1|\|_2 \leq q(n)\,\mathtt{u} \left\| |\widehat{L}| \left[|\widehat{L}^{-1}||\widehat{L}||\widehat{D}||\widehat{L}^T||\widehat{L}^{-T}| \right]_{[L+\frac{1}{2}D]} |\widehat{D}^{-1}||\widehat{L}^{-1}| \right\|_2. \tag{4.3.21}$$

Si definimos $I_n + E = (I_n + E_3)(I_n + E_2)(I_n + P^T E_1 P)$, entonces (4.3.19) se convierte en (4.3.16), donde

$$\|E\|_2 \leq \|E_1\|_2 + \|E_2\|_2 + \|E_3\|_2 \leq \||\,|E_1|\,\|_2 + \|E_2\|_2 + \|E_3\|_2. \tag{4.3.22}$$

Si en (4.3.22), reemplazamos (4.3.18),(4.3.20) y (4.3.21) tenemos (4.3.17). □

El análisis de errores presentado en el Teorema 4.3.4 nos permite expresar el error regresivo cometido por el Algoritmo 4.3.1 como una perturbación multiplicativa de la matriz. Combinando este resultado con los Teoremas 2.2.2 y 2.2.3 obtenemos nuestro resultado principal, el cual establece la precisión obtenida para los autovalores y autovectores calculados por el Algoritmo 4.3.1.

Corolario 4.3.5 *Sea $A \in \mathbb{R}^{n\times n}$ una matriz simétrica y no singular, y $S = diag[s_1, \ldots, s_n]$ con $s_i > 0$, y sean $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_n$ y $\widehat{q}_1, \ldots, \widehat{q}_n$, respectivamente, los autovalores y los autovectores calculados por el Algoritmo 4.3.1 en un ordenador con unidad de redondeo* $\mathtt{u}$. *Ahora, supongamos que $\lambda_1, \ldots, \lambda_n$ y $q_1, \ldots, q_n$ son respectivamente, los autovalores y los autovectores ortogonales exactos de A. Entonces*

$$\frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq q(n)\,\mathtt{u}\left(\Xi + \kappa_2(\widehat{L})\right) + O(\mathtt{u}^2)$$
$$\sin\theta(q_i, \widehat{q}_i) \leq q(n)\,\mathtt{u}\left(\Xi + \kappa_2(\widehat{L})\right)\left(1 + \frac{2}{relgap_{\widehat{\lambda}}(A, \lambda_i)}\right) + O(\mathtt{u}^2),$$

donde

$$\Xi := \left\| |\widehat{L}|\left[|\widehat{L}^{-1}||\widehat{L}||\widehat{D}||\widehat{L}^T||\widehat{L}^{-T}|\right]_{[L+\frac{1}{2}D]}|\widehat{D}^{-1}||\widehat{L}^{-1}| \right\|_2,$$

$q(n)$ es un polinomio cuadrático en n y $\theta(q_i, \widehat{q}_i)$ es el ángulo agudo entre q_i y $\widehat{q}_i$.

El Corolario 4.3.5 proporciona cotas para el error progresivo en los autovalores y en los autovectores calculados por el Algoritmo 4.3.1 en términos de cantidas calculadas. Sin embargo, si como en el Teorema 4.3.2, disponemos de alguna información adicional relacionada con el escalamiento de la matriz, la precisión obtenida de los autovalores y autovectores calculados por el Algoritmo 4.3.1 se puede escribir dependiendo explícitamente del escalamiento:

Corolario 4.3.6 *Con las mismas hipótesis del Corolario 4.3.5 y supongamos además que, para cualquier matriz $S = diag[s_1, \ldots, s_n]$ con $s_i > 0$, definimos*

$\widehat{L}_B := S^{-1}\widehat{L}S$, $\widehat{D}_B := S^{-1}\widehat{D}S^{-1}$, *entonces*

$$\frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq q(n)\,\mathtt{u}\left(\tau\,\Xi_B + \kappa_2(\widehat{L})\right) + O(\mathtt{u}^2) \tag{4.3.23}$$

$$\sin\theta(q_i, \widehat{q}_i) \leq q(n)\,\mathtt{u}\left(\tau\,\Xi_B + \kappa_2(\widehat{L})\right)\left(1 + \frac{2}{relgap_{\widehat{\lambda}}(A, \lambda_i)}\right) + O(\mathtt{u}^2) \tag{4.3.24}$$

donde

$$\Xi_B := \left\| |\widehat{L}_B| \left[|\widehat{L}_B^{-1}||\widehat{L}_B||\widehat{D}_B||\widehat{L}_B^T||\widehat{L}_B^{-T}| \right]_{[L+\frac{1}{2}D]} |\widehat{D}_B^{-1}||\widehat{L}_B^{-1}| \right\|_2, \tag{4.3.25}$$

y

$$\tau := max\{1, \tau_L, \tau_D\}, \quad \tau_L := \max_{j<k} \frac{s_k}{s_j} \quad y \quad \tau_D := \max_{blocks,i} \max \left\{ \frac{s_{i+1}}{s_i}, \frac{s_i}{s_{i+1}} \right\}.$$

Concluimos que el Algoritmo 4.3.1 calculará de manera precisa *todos* los autovalores con el signo y número de dígitos principales correctos y los autovectores correspondientes a los autovalores bien separados de una matriz graduada A, siempre y cuando se cumplan las siguientes condiciones:

1. Existe un escalamiento S de la matriz A tal que la factorización de Bunch & Parlett

   $$PAP^T \simeq \widehat{L}\widehat{D}\widehat{L}^T = S\,S^{-1}\widehat{L}S\,S^{-1}\widehat{D}S^{-1}\,S\widehat{L}^T S^{-1}\,S := S\widehat{L}_B\widehat{D}_B\widehat{L}_B^T S$$

   proporciona matrices $\widehat{L}_B$ y $\widehat{D}_B$ con números de condición moderados.

2. Los factores del escalamiento τ_L y τ_D son pequeños. Si PAP^T está bien escalada en el sentido usual, con los elementos en la diagonal de S ordenados decrecientemente, entonces $\tau_L \leq 1$, pero esto sólamente no garantiza alta precisión para una matriz graduada simétrica indefinida. También es necesario que el *nuevo* factor τ_D sea pequeño. El factor τ_D proviene de la presencia de los bloques 2×2 en la factorización de Bunch & Parlett de A. Si existe un bloque 2×2 en la posición i, $i+1$, de la matriz diagonal por bloques D, y un "salto" ya sea creciente o decreciente, en los elementos diagonales de S, s_i y s_{i+1}, habrá un "condicionamiento efectivo" de tamaño τ_D que amplifica las perturbaciones de entrada de orden $\mathtt{u}$.

Insistimos una vez más que la principal novedad en este capítulo es la presencia del factor τ_D en las cotas para la alta precisión de los autovalores y autovectores de matrices graduadas simétricas indefinidas. Esto se ha demostrado en el Corolario 4.3.6. En la siguiente sección mostraremos algunos experimentos numéricos para confirmar la presencia de este factor.

4.4. Experimentos y ejemplos numérico

Hemos demostrado rigurosamente en el Corolario 4.3.6 que el Algoritmo 4.3.1 calcula los autovalores y autovectores de una matriz simétrica graduada A con alta precisión relativa siempre que A esté bien escalada y que los números de condición de las matrices escaladas $\widehat{L}_B$ y $\widehat{D}_B$ sean pequeños. En esta sección presentaremos un experimento numérico para comprobar la dependencia de τ en los errores relativos progresivos de los autovalores y autovectores de matrices simétricas graduadas. Comparamos respectivamente, los autovalores y los autovectores calculados por el Algoritmo 4.3.1, $\{\widehat{\lambda}_i\}_{i=1}^n$ y $\{\widehat{q}_i\}_{i=1}^n$, con los autovalores y los autovectores "exactos", $\{\lambda_i\}_{i=1}^n$ y $\{q_i\}_{i=1}^n$, calculados con la función `eig` de MATLAB™ con 100 dígitos de precisión. Además, presentamos un ejemplo numérico para ilustrar que los resultados en el Corolario 4.3.6 son intrínsecos, es decir, la presencia del factor τ_D es independiente del algoritmo utilizado para el cálculo. Todos los experimentos en esta sección han sido desarrollados en MATLAB R2012a con $\mathtt{u} = 2^{-53}$. A partir de las ecuaciones (4.3.23) y (4.3.24) en el Corolario 4.3.6 para los errores progresivos en los autovalores y en los autovectores tenemos que:

$$\frac{1}{\mathtt{u}\,\Xi_B}\frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} \lesssim C_1\,\tau \tag{4.4.1}$$

y

$$\frac{1}{\mathtt{u}\,\Xi_B}\frac{\sin\theta(q_i,\widehat{q}_i)}{\left(1+\frac{2}{relgap_{\widehat{\lambda}}(A,\lambda_i)}\right)} \lesssim C_2\,\tau, \tag{4.4.2}$$

para algunas constantes C_1 y C_2. En cada experimento evaluamos numéricamente las expresiones

$$\Phi = \frac{1}{\mathtt{u}\,\Xi_B}\max_i \frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} \tag{4.4.3}$$

y

$$\Psi = \frac{1}{\mathtt{u}\,\Xi_B}\frac{\max_i \sin\theta(q_i,\widehat{q}_i)}{\left(1+\frac{2}{relgap_{\widehat{\lambda}}(A,\lambda_i)}\right)}. \tag{4.4.4}$$

En nuestros experimentos consideramos que $\Xi_B \approx \kappa^3(L_B)\kappa(D_B)$, por lo tanto, esperamos que Φ y Ψ se comporten como múltiplos de τ, τcomo en el Corolario 4.3.6.

4.4.1. Experimento numérico

En este experimento numérico definimos una estructura por bloques para la factorización BP-LDL^T de $B = L_0 D_0 L_0^T$ con al menos un bloque 2×2. Ahora, generamos una matriz diagonal $\underline{S} \in \mathbb{R}^{n \times n}$ tal que sus elementos diagonales son potencias de diez, finalmente generamos una matriz de permutaciones aleatorias P_0 y con esto construimos la matriz simétrica graduada $A = P_0 \underline{S} L_B D_B L_B^T \underline{S} P_0^T$. El Algoritmo 4.3.1 es aplicado a estas matrices para calcular los autovalores. En el Paso 1 se lleva a cabo la factorización por bloques de A: $PAP^T \approx \widehat{L}\widehat{D}\widehat{L}^T$, con esto definimos $S = P\underline{S}P^T, \widehat{L}_B = S^{-1}\widehat{L}S$ y $\widehat{D}_B = S^{-1}\widehat{D}S^{-1}$. El factor τ_L siempre es menor o igual que uno, ya que la factorización de Bunch & Parlett utiliza pivote completo.

Para este proceso generamos matrices aleatorias utilizando los comandos `rand` y `randn` de MATLAB™ y consideramos diferentes tamaños de matrices: 50×50, 100×100 y 200×200. Las entradas diagonales de $\underline{S}$ son potencias enteras de diez con los exponentes elegidos de dos maneras. En la primera forma de elección de los exponentes, elegimos aleatoriamente las potencias en el intervalo [-10, 10], y en la segunda, en las posiciones correspondientes a uno de los bloques 2×2 los elementos diagonales de $\underline{S}$ son $\underline{S}(i,i) = \tau_D \, 10^{s_i}$ y $\underline{S}(i+1, i+1) = 10^{s_i}$ donde s_i es un número entero aleatorio que pertenece al mismo intervalo del resto de los exponentes.

Para cada valor de τ, consideramos 300 muestras y calculamos el valor máximo de Φ, para el caso de los autovalores y el valor máximo de Ψ, para los autovectores. Obtenemos resultados similares en todos los casos. En las Figuras 4.4.1 y 4.4.2 hemos graficado respectivamente, los resultados para los errores relativos de los autovalores y de los autovectores para las matrices simétricas graduadas de orden 100 descritas anteriormente.

Puede observarse que el factor τ aparece en el máximo error relativo de los autovalores y también en el máximo error relativo de los autovectores como lo demostramos en el Corolario 4.3.6.

4.4.2. Ejemplo numérico

Presentamos ahora un ejemplo numérico, con el cual podemos observar la sensibilidad del problema de autovalores de una matriz graduada simétrica independien-

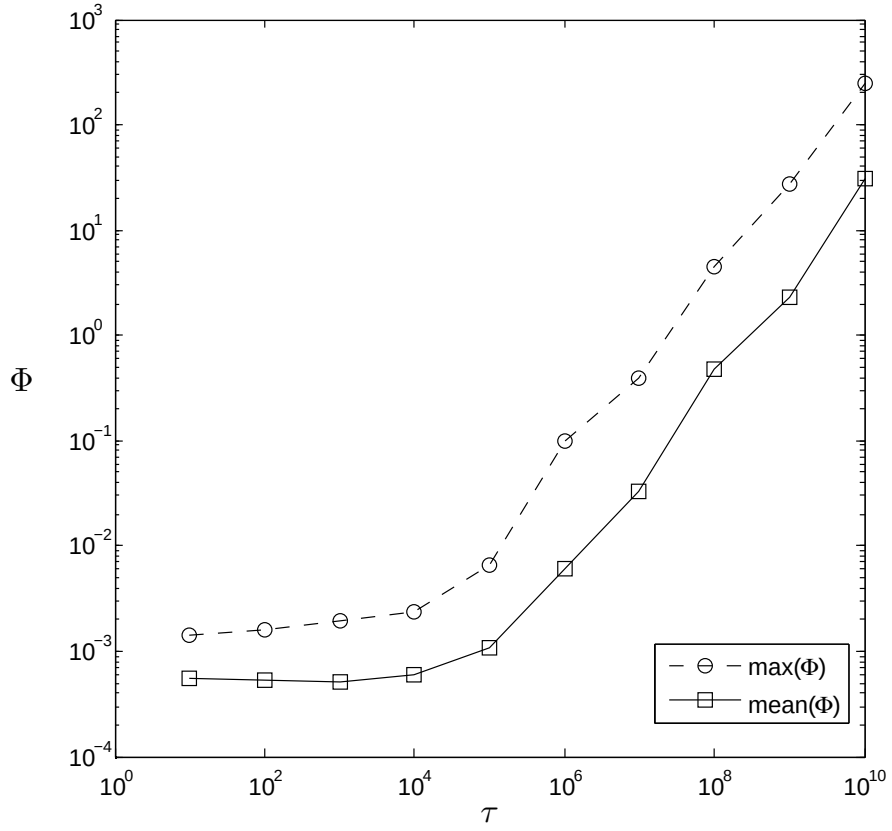

Figura 4.4.1: Error relativo progresivo en los autovalores. Φ como en la ecuación 4.4.3 frente τ para matrices simétricas graduadas aleatorias $A = SBS$ con $B = L_0 D_0 L_0^T$ de orden $n = 100$, con al menos un bloque 2×2.

temente del algoritmo usado. Sea $A = SBS$ la siguiente matriz graduada

$$A = \begin{bmatrix} 3{,}00e+20 & 1{,}50e+20 & 1{,}50e+10 & -1{,}50e+10 \\ 1{,}50e+20 & 7{,}50e+19 & 2{,}25e+10 & -2{,}50e+09 \\ 1{,}50e+10 & 2{,}25e+10 & 7{,}5e-01 & -6{,}00e-01 \\ -1{,}50e+10 & -2{,}50e+09 & -6{,}00e-01 & 1{,}35e+00 \end{bmatrix}$$

donde

$$S = \begin{bmatrix} 1{,}0e+10 & 0 & 0 & 0 \\ 0 & 1{,}0e+10 & 0 & 0 \\ 0 & 0 & 1{,}0e+00 & 0 \\ 0 & 0 & 0 & 1{,}0e+00 \end{bmatrix}$$

y

$$B = \begin{bmatrix} 3{,}0 & 1{,}50 & 1{,}50 & -1{,}50 \\ 1{,}5 & 0{,}75 & 2{,}25 & -0{,}25 \\ 1{,}5 & 2{,}25 & 0{,}75 & -0{,}60 \\ -1{,}5 & -0{,}25 & -0{,}60 & 1{,}35 \end{bmatrix}.$$

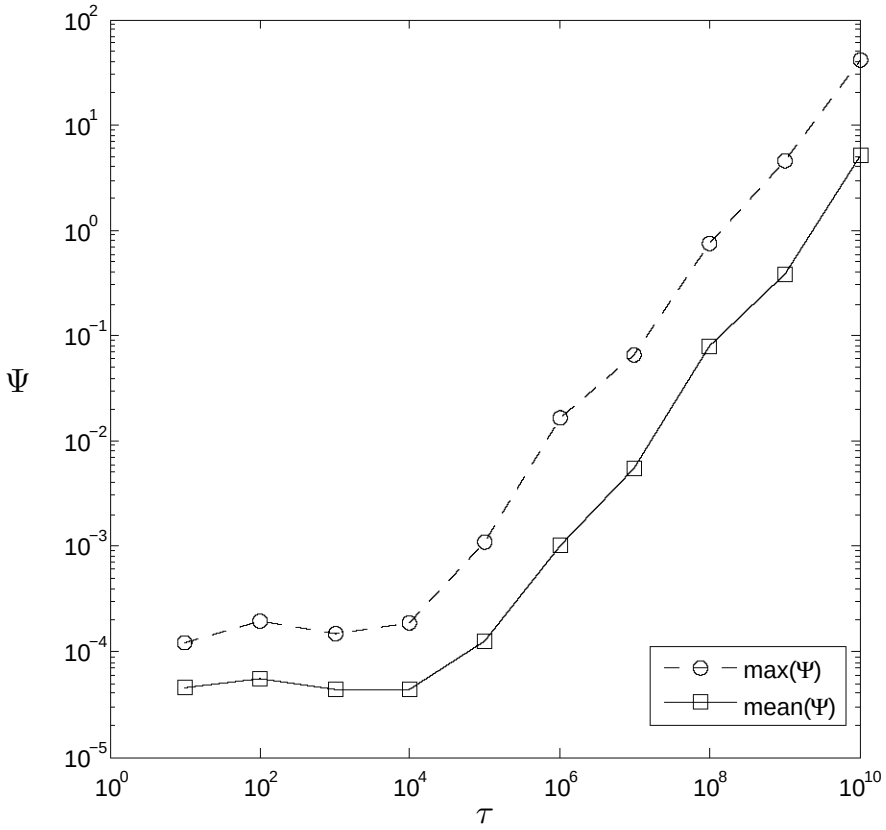

Figura 4.4.2: Error relatico progresivo en los autovectores. Ψ como en la ecuación 4.4.4 frente τ para matrices simétricas graduadas aleatorias $A = SBS$ con $B = L_0 D_0 L_0^T$ de orden $n = 100$, con al menos un bloque 2×2.

Sea $\widetilde{A} = S\widetilde{B}S$ una perturbación de la matriz A con $\widetilde{B} = B + \delta B$, donde δB es una matriz de perturbaciones simétrica tal que $\|\delta B\| = 10^{-12}\|B\|$.

La matriz A está muy mal condicionada, ya que $\kappa(A) = 8{,}33e + 20$, pero A es una matriz graduada ya que $\kappa(B) = 21{,}06$.

Ahora, calculamos una factorización por bloques BP-LDL^T de A:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0{,}5 & 1 & 0 & 0 \\ 5e-11 & 0 & 1 & 0 \\ -5e-11 & 1e-11 & 0{,}33 & 1 \end{bmatrix}$$

y

$$D = \begin{bmatrix} 3e+20 & 0 & 0 & 0 \\ 0 & 0 & 1{,}5e+10 & 0 \\ 0 & 1{,}5e+10 & 0 & 0 \\ 0 & 0 & 0 & 0{,}5 \end{bmatrix},$$

con $\kappa(D) = 5{,}99e20$ y $\kappa(L) = 1{,}64$. Sean

$$L_B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0{,}5 & 1 & 0 & 0 \\ 0{,}5 & 0 & 1 & 0 \\ -0{,}5 & 0{,}1 & 0{,}33 & 1 \end{bmatrix} \quad \text{y} \quad D_B = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 0 & 1{,}5 & 0 \\ 0 & 1{,}5 & 0 & 0 \\ 0 & 0 & 0 & 0{,}5 \end{bmatrix}$$

los factores de la factorización por bloques BP-LDL^T de B, tales que $\kappa_2(D_B) = 5{,}99$ y $\kappa_2(L_B) = 2{,}51$. La relación entre todas estas matrices es

$$A = SBS = SL_BD_BL_B^TS = (SL_BS^{-1})(SD_BS)(S^{-1}L_B^TS) = L_AD_AL_A^T.$$

Los autovalores "exactos" de A y $\widetilde{A}$ calculados con la función `eig` de MATLAB™ con 32 dígitos son respectivamente:

$$\lambda = \begin{bmatrix} -14142135623{,}922617154686463358538 \\ 0{,}449999999999999999974575 \\ 14142135623{,}539283821353130025209 \\ 37500000000000000002{,}03333333333 \end{bmatrix}$$

y

$$\widetilde{\lambda} = \begin{bmatrix} -14093173224{,}088992490682477403928 \\ 0{,}450000000001101213712274596 6807 \\ 14191268128{,}178437675970947260619 \\ 37500000000249075736{,}18576490472 \end{bmatrix}.$$

Según la ecuación (4.2.21), en el Corolario 4.2.4, la cota, a primer orden en ϵ, para el máximo error relativo entre los autovalores de $A = SBS$ y $\widetilde{A} = S(B + \delta B)S$, con $\|\delta B\| \leq \epsilon \|B\|$, está dada por:

$$\frac{|\widetilde{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq 2\,\epsilon\,\tau\,\kappa_2^3(L_B)\,\kappa_2(D_B) \approx 2 \cdot\ 10^{-9}\,\tau.$$

En este ejemplo, y sin considerar los factores pesimistas $\kappa_2^3(L_B)\,\kappa_2(D_B)$ pero tomando en cuenta que $\tau = \tau_D = 10^{10}$ y $\epsilon = 10^{-12}$, obtendríamos que

$$\max_i \frac{|\widetilde{\lambda}_i - \lambda_i|}{|\lambda_i|} \lesssim\ 10^{-2}. \tag{4.4.5}$$

Las diferencias relativas entre los autovalores "exactos" de A y $\widetilde{A}$ son las siguientes:

$$\frac{|\widetilde{\lambda}_i - \lambda_i|}{|\lambda_i|} \simeq \begin{bmatrix} 3{,}46e-03 \\ 2{,}45e-13 \\ 3{,}47e-03 \\ 6{,}64e-13 \end{bmatrix}.$$

Observamos que el primer y el tercer error relativo en los autovalores son de orden $\epsilon\tau$, y que el segundo y el cuarto son de orden ϵ. El factor τ aparece tal y como lo indicamos en la ecuación (4.4.5).

Según la ecuación (4.2.22) en el Corolario 4.2.4, la cota, a primer orden en ϵ, para el seno del ángulo entre los autovectores de $A = SBS$ y $\widetilde{A} = S(B + \delta B)S$, con $\|\delta B\| \leq \epsilon \|B\|$, está dada por:

$$\sin\theta(q_i, \widetilde{q}_i) \leq \epsilon\, \tau\, \kappa_2^3(L_B)\, \kappa_2(D_B) \left(1 + \frac{2}{relgap_{\widetilde{\lambda}}(A, \lambda)}\right). \tag{4.4.6}$$

Procediendo como en la ecuación (4.4.5), esperamos que el factor τ también aparezca en el error relativo de los autovectores, es decir

$$\max_i(\sin\theta(q_i, \widetilde{q}_i)) \lesssim 10^{-2}. \tag{4.4.7}$$

Los errores relativos entre los autovectores "exactos" de A y $\widehat{A}$ son:

$$\sin\theta(q_i, \widetilde{q}_i) \simeq \begin{bmatrix} 1{,}73e-03 \\ 8{,}06e-13 \\ 1{,}73e-03 \\ 4{,}78e-13 \end{bmatrix},$$

de la misma manera que para las diferencias relativas de los autovalores, observamos que el primer y el tercer error relativo en los autovectores son de orden $\epsilon\tau$, y que el segundo y el cuarto son de orden ϵ.

5

Conclusiones y trabajos futuros

La idea clave de esta memoria es la obtención de Alta Precisión Relativa (HRA) para cálculos en Álgebra Lineal Numérica. Esto se hace mediante algoritmos que usan Descomposiciones que Revelan el Rango (Rank Revealing Decompositions, RRDs) precisas y que luego aprovechan la estructura específica del problema. Este esquema ya se había aplicado con éxito a Problemas de Valores Singulares (SVD), al Problema Simétrico de Autovalores y Autovectores (SEVP) y a Sistemas de Ecuaciones Lineales (LSE). Nosotros lo hemos extendido a Problemas de Mínimos Cuadrados (Capítulo 3) y al Problema Simétrico de Autovalores y Autovectores para un tipo especial de matrices: Graduadas Simétricas (Capítulo 4).

En los Capítulos 3 y 4 hemos presentado y analizado cuidadosamente dos nuevos algoritmos, para calcular respectivamente soluciones precisas para problemas de mínimos cuadrados $\min_{x\in\mathbb{R}^n} \|Ax-b\|_2$, y soluciones precisas para problemas de autovalores y autovectores de matrices simétricas graduadas, tales que una descomposición que revele el rango precisa de la matriz de coeficientes A pueda ser calculada. Como lo explicamos en la introducción del Capítulo 2, esto es posible para diferentes clases de matrices estructuradas que pueden tener números de condición extremadamente grandes, y probablemente, será posible para más clases en el futuro. Adicionalmente, el Algoritmo 3.5.1 también puede aplicarse para calcular de manera precisa soluciones mínimas de sistemas lineales infradeterminados.

Los trabajos existentes en el campo al comienzo de mi investigación han dado la pauta, pero en este memoria se presentan nuevos algoritmos, Algoritmo 3.5.1 y 4.3.1; nuevos análisis de errores, Teorema 3.5.2 y Corolario 4.3.5; y nuevos resultados de teoría de perturbaciones de la pseudo inversa de Moore-Penrose, Teoremas 3.2.2, 3.2.3 y 3.2.5, nuevos resultados de teoría de perturbaciones de Problemas de

Mínimos Cuadrados 3.3.1 y nuevos resultados de teoría de perturbaciones de Problema Simétrico de Autovalores y Autovectores para Matrices Graduadas Simétricas, Corolario 4.2.4.

Esta investigación ha dado lugar a tres publicaciones: el contenido del Capítulo 3 corresponde a los artículos [11] y [12]. El artículo [11], ha sido aceptado recientemente por la revista *SIAM Journal on Matrix Analysis*, y el artículo [12] se someterá en breve a la revista *Linear Algebra and its Applications*. El contenido del Capítulo 4 corresponde al artículo [13] que ha sido enviado a la revista *Electronic Transactions on Numerical Analysis*.

Además la investigación incluida en esta tesis ha sido presentada en los siguientes Congresos Internacionales, los describiremos haciendo referencia al capítulo correspondiente.

El contenido del Capítulo 3, se ha presentado en:

- Householder Symposium XVIII. Tahoe City - Estados Unidos. Junio 2011.
- 7th International Congress on Industrial and Applied Mathematics. Vancouver - Canadá. Julio 2011.
- The 17th International Linear Algebra Society Conference. Braunschweig - Alemania. Agosto 2011.
- 2012 SIAM Conference on Applied Linear Algebra. Valencia - España. Junio 2012.

El contenido del Capítulo 4, se ha presentado en:

- XVII Congreso Colombiano de Matemáticas. Cali - Colombia. Agosto de 2009.
- Álgebra Lineal, Análisis Matricial y Aplicaciones - ALAMA 2012. Leganés - España. Junio 2012.

En la actualidad estamos trabajando en calcular la SVD con HRA usando un algoritmo de tipo Jacobi implícito en las líneas presentadas en [30], este trabajo se ha presentado en:

- VIII International Workshop on Accurate Solution of Eigenvalue Problems. Berlín - Alemania. Junio 2010.

- The 16th ILAS Conference of the International Linear Algebra Society. Pisa - Italia. Junio 2010.
- XVIII Congreso Colombiano de Matemáticas. Bucaramanga - Colombia. Julio 2011.

Esta tesis junto con otros trabajos existentes [20, 30, 32] muestran que, para aquellas matrices para las cuales puede calcularse una descomposición que revele el rango precisa, podemos desarrollar de manera precisa y eficiente casi todas las tareas clásicas del *Álgebra Lineal Numérica*, es decir, solución de sistemas lineales, solución de problemas de mínimos cuadados, cálculo de autovalores y autovectores de matrices simétricas y descomposición de valores singulares, para obtener errores relativos de orden $\mathtt{u}$ para problemas muy mal condicionados donde algoritmos estándar fallan en proporcionar hasta un dígito correcto de precisión. El único problema básico que es excluido de este marco teórico es el problema no simétrico de autovalores. Investigar hasta donde la descomposición que revela el rango nos permite resolver de manera precisa problemas no simétricos de autovalores será el objetivo de nuestra futura investigación.

Bibliografía

[1] E. Anderson, Z. Bai, C. Bischof, L. S. Blackford, J. Demmel, Jack J. Dongarra, J. Du Croz, S. Hammarling, A. Greenbaum, A. McKenney, and D. Sorensen. *LAPACK Users' guide (third ed.)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999.

[2] J. M. Banoczi, N.-C. Chiu, G. E. Cho, and I. C. F. Ipsen. The lack of influence of the right-hand side on the accuracy of linear system solution. *SIAM Journal on Scientific Computing*, 20(1):203–227, 1998.

[3] J. Barlow and J. Demmel. Computing accurate eigensystems of scaled diagonally dominant matrices. *SIAM Journal on Numerical Analysis*, 27(3):762–791, June 1990.

[4] J.L. Barlow. Error analysis and implementation aspects of deferred correction for equality constrained least squares problems. *SIAM Journal on Numerical Analysis*, 25:1340–1358, 1988.

[5] Å. Björck. *Numerical methods for least squares problems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.

[6] Å. Björck and V. Pereyra. Solution of Vandermonde systems of equations. *Mathematics of Computation*, 24:893–903, 1970.

[7] J. R. Bunch. Analysis of the diagonal pivoting method. *SIAM Journal on Numerical Analysis*, 8(4):656–680, 1971.

[8] J. R. Bunch and B. N. Parlett. Direct methods for solving symmetric indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 8:639–655, 1971.

[9] L.-X. Cai, W.-W. Xu, and W. Li. Additive and multiplicative perturbation bounds for the Moore-Penrose inverse. *Linear Algebra and its Applications*, 434(2):480–489, 2011.

[10] S. L. Campbell and C. D. Meyer, Jr. *Generalized inverses of linear transformations*. Dover Publications Inc., New York, 1991. Corrected reprint of the 1979 original.

[11] N. Castro-González, J. Ceballos, F. M. Dopico, and J. M Molera. Accurate solution of structured least squares problems via rank-revealing decompositions. *SIAM Journal on Matrix Analysis and Applications*, 2012. Accepted for publication.

[12] N. Castro-González, J. Ceballos, F. M. Dopico, and J. M. Molera. Multiplicative perturbation theory of the Moore-Penrose inverse and the least squares problem. *in preparation*, 2013.

[13] J. Ceballos, F. M. Dopico, and J. M. Molera. Computing accurate eigenvalues and eigenvectors of graded symmetric indefinite matrices. *Submitted to be published in Electronic Transactions on Numerical Analysis*, 2013.

[14] T. F. Chan and D. E. Foulser. Effectively well-conditioned linear systems. *SIAM Journal on Scientific and Statistical Computing*, 9(6):963–969, 1988.

[15] B. A. Cipra. The Best of the 20th Century: Editors Name Top 10 Algorithms. *SIAM News*, 33(4), 2000.

[16] A. J. Cox and N. J. Higham. Stability of Householder QR factorization for weighted least squares problems. In D.F. Griffiths, D.J. Higham, and G.A. Watson, editors, *Numerical Analysis 1997, Proceedings of the 17th Dundee Conference*, pages 57–73, Harlow, Essex, UK, 1998. Addison-Wesley-Longman.

[17] J. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, 1997.

[18] J. Demmel. Accurate singular value decompositions of structured matrices. *SIAM Journal on Matrix Analysis and Applications*, 21(2):562–580, 1999.

[19] J. Demmel and W. Gragg. On computing accurate singular values and eigenvalues of acyclic matrices. *Linear Algebra and its Applications*, 185:203–218, 1993.

[20] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač. Computing the singular value decomposition with high relative accuracy. *Linear Algebra and its Applications*, 299(1–3):21–80, 1999.

[21] J. Demmel and W. Kahan. Accurate singular values of bidiagonal matrices. *SIAM Journal on Scientific and Statistical Computing*, 11(5):873–912, September 1990.

[22] J. Demmel and P. Koev. Necessary and sufficient conditions for accurate and efficient rational function evaluation and factorizations of rational matrices. In *Structured matrices in mathematics, computer science, and engineering, II (Boulder, CO, 1999)*, volume 281 of *Contemporary Mathematics*, pages 117–143. American Mathematical Society, Providence, RI, 2001.

[23] J. Demmel and P. Koev. Accurate SVDs of weakly diagonally dominant M-matrices. *Numerische Mathematik*, 98(1):99–104, 2004.

[24] J. Demmel and P. Koev. Accurate SVDs of polynomial Vandermonde matrices involving orthonormal polynomials. *Linear Algebra and its Applications*, 417(2-3):382–396, 2006.

[25] J. Demmel and K. Veselić. Jacobi's method is more accurate than QR. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1204–1246, 1992.

[26] I. S. Dhillon and B. N Parlett. Relatively robust representations of symmetric tridiagonals. *Linear Algebra and its Applications*, 309(1–3):121 – 151, 2000.

[27] I. S. Dhillon and B. N. Parlett. Orthogonal eigenvectors and relative gaps. *SIAM Journal on Matrix Analysis and Applications*, 25(3):858–899, 2004.

[28] F. M. Dopico and P. Koev. Accurate symmetric rank revealing and eigendecompositions of symmetric structured matrices. *SIAM Journal on Matrix Analysis and Applications*, 28(4):1126–1156, 2006.

[29] F. M. Dopico and P. Koev. Perturbation theory for the LDU factorization and accurate computations for diagonally dominant matrices. *Numerische Mathematik*, 119(2):337–371, 2011.

[30] F. M. Dopico, P. Koev, and J. M. Molera. Implicit standard Jacobi gives high relative accuracy. *Numerische Mathematik*, 113(2):519–553, 2009.

[31] F. M. Dopico and J. M Molera. Perturbation theory for factorizations of LU type through series expansions. *SIAM Journal on Matrix Analysis and Applications*, 27(2):561–581, 2005.

[32] F. M. Dopico and J. M. Molera. Accurate solution of structured linear systems via rank-revealing decompositions. *IMA Journal of Numerical Analysis*, 32:1096–1116, 2012.

[33] F. M. Dopico, J. M. Molera, and J. Moro. An orthogonal high relative accuracy algorithm for the symmetric eigenproblem. *SIAM Journal on Matrix Analysis and Applications*, 25(2):301–351, 2003.

[34] Z. Drmač. Accurate computation of the product induced singular value decomposition with applications. *SIAM Journal of Numerical Analysis,*, 35(5):1969–1994, 1998.

[35] Z. Drmač and K. Veselić. New fast and accurate Jacobi SVD algorithm. I. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1322–1342, 2008.

[36] Z. Drmač and K. Veselić. New fast and accurate Jacobi SVD algorithm. II. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1343–1362, 2008.

[37] S. Eisenstat and I. Ipsen. Relative perturbation techniques for singular value problems. *SIAM Journal on Numerical Analysis*, 32(6):1972–1988, 1995.

[38] K. Fernando and B. Parlett. Accurate singular values and differential qd algorithms. *Numerische Mathematik*, 67:191–229, 1994.

[39] G. Golub and C. Van Loan. *Matrix computations*. Johns Hopkins University Press, 4th edition, 2012.

[40] M. Gu and S. C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

[41] N. J. Higham. The Test Matrix Toolbox for Matlab (version 3.0). *Numerical Analysis Report No. 276, Manchester Center for Computational Mathematics, Manchester, England*, 1995.

[42] N. J. Higham. QR factorization with complete pivoting and accurate computation of the SVD. *Linear Algebra and its Applications*, 309(1-3):153–174, 2000.

[43] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2002.

[44] P. D. Hough and S. A. Vavasis. Complete orthogonal decomposition for weighted least squares. *SIAM Journal on Matrix Analysis and Applications*, 18(2):369–392, 1997.

[45] I. C. F. Ipsen. Relative perturbation results for matrix eigenvalues and singular values. In *Acta numerica, 1998*, volume 7 of *Acta Numer.*, pages 151–201. Cambridge Univ. Press, Cambridge, 1998.

[46] I. C. F. Ipsen. An overview of relative $\sin\Theta$ theorems for invariant subspaces of complex matrices. *J. Comput. Appl. Math.*, 123(1-2):131–153, 2000. Numerical analysis 2000, Vol. III. Linear algebra.

[47] W. Kahan. Accurate eigenvalues of a symmetric tridiagonal matrix. Computer Science Dept. Technical Report CS41, Stanford University, Stanford, CA, July 1966 (revised June 1968).

[48] P. Koev. Accurate eigenvalues and SVDs of totally nonnegative matrices. *SIAM Journal on Matrix Analysis and Applications*, 27(1):1–23, 2005.

[49] D. Kressner. *Numerical methods for general and structured eigenvalue problems*, volume 46 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin, 2005.

[50] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users' Guide*. Society for Industrial and Applied Mathematics, January 1998.

[51] R-C. Li. Relative perturbation theory. I. Eigenvalue and singular value variations. *SIAM Journal on Matrix Analysis and Applications*, 19(4):956–982, 1998.

[52] R-C. Li. Relative perturbation theory. II. Eigenspace and singular subspace variations. *SIAM Journal on Matrix Analysis and Applications*, 20(2):471–492, 1999.

[53] A. Marco and J. J. Martínez. Polynomial least squares fitting in the Bernstein basis. *Linear Algebra and its Applications*, 433(7):1254–1264, 2010.

[54] R. Martin, C. Reinsch, and J. H. Wilkinson. Householder tridiagonalization of a real symmetric matrix. *Numerische Mathematik*, 11:181–195, 1968.

[55] R. Mathias. Accurate eigensystem computations by Jacobi methods. *SIAM Journal on Matrix Analysis and Applications*, 16(3):977–1003, 1995.

[56] R. Mathias. Spectral perturbation bounds for positive definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 18(4):959–980, 1997.

[57] L. Miranian and M. Gu. Strong rank revealing *LU* factorizations. *Linear Algebra and its Applications*, 367:1–16, 2003.

[58] V. Olshevsky, editor. *Structured matrices in mathematics, computer science, and engineering. I*, volume 280 of *Contemporary Mathematics*, Providence, RI, 2001. American Mathematical Society.

[59] V. Olshevsky, editor. *Structured matrices in mathematics, computer science, and engineering. II*, volume 281 of *Contemporary Mathematics*, Providence, RI, 2001. American Mathematical Society.

[60] C.-T. Pan. On the existence and computation of rank-revealing *LU* factorizations. *Linear Algebra and its Applications*, 316(1-3):199–222, 2000.

[61] B. N. Parlett. *The symmetric eigenvalue problem*. SIAM, Philadelphia, 1998.

[62] M. J. Peláez and J. Moro. Accurate factorization and eigenvalue algorithms for symmetric DSTU and TSC matrices. *SIAM Journal on Matrix Analysis and Applications*, 28(4):1173–1198, 2006.

[63] J. M. Peña. LDU decompositions with L and U well conditioned. *Electronic Transactions on Numerical Analysis*, 18:198–208 (electronic), 2004.

[64] M.J.D. Powell and J.K. Reid. On applying Householder transformations to linear least squares problems. In *Information Processing 68, Proc. International Federation of Information Processing Congress, Edinburgh, 1968*, pages 122–126. North Holland, Amsterdam, 1969.

[65] I. Slapničar. Componentwise analysis of direct factorization of real symmetric and Hermitian matrices. *Linear Algebra and its Applications*, 272:227–275, 1998.

[66] I. Slapničar. Highly accurate symmetric eigenvalue decomposition and hyperbolic SVD. *Linear Algebra and its Applications*, 358:387–424, 2003.

[67] I. Slapničar. *Accurate symmetric eigenreduction by a Jacobi method*. PhD thesis, Fernuniversität - Hagen, Hagen, Germany, 1992.

[68] G. W. Stewart. On the perturbation of LU, Cholesky, and QR factorizations. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1141–1145, 1993.

[69] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, New York, 1990.

[70] K. Veselić. A Jacobi eigenreduction algorithm for definite matrix pairs. *Numerische Mathematik*, 64:241–269, 1993.

[71] K. Veselić and V. Hari. A note on a one-sided jacobi algorithm. *Numerische Mathematik*, 56(6):627–633, 1989.

[72] K. Veselić and I. Slapničar. Floating point perturbations of Hermitian matrices. *Linear Algebra and its Applications*, 195:81–116, 1993.

[73] D. S. Watkins. *The matrix eigenvalue problem: GR and Krylov subspace methods*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

[74] P.-Å. Wedin. Perturbation theory for pseudo-inverses. *BIT*, 13:217–232, 1973.

[75] Q. Ye. Computing singular values of diagonally dominant matrices to high relative accuracy. *Mathematics of Computation*, 77(264):2195–2230, 2008.